Studies of evolution from a genomic perspective

by

Hunter Bryan Fraser

B.S. (Massachusetts Institute of Technology) 2001

A dissertation filed in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge,

Professor Michael B. Eisen, Chair

Professor Nicholas Cozzarelli

Professor Jasper Rine

Professor Lior Pachter

Spring 2005

The dissertation of Hunter Bryan Fraser is approved,

_____

Chair Date

_____

Date

_____

Date

_____

Date

University of California, Berkeley

Spring 2005

Studies of evolution from a genomic perspective

by

Hunter Bryan Fraser

ABSTRACT

Studies of evolution from a genomic perspective

by

Hunter Bryan Fraser

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Michael B. Eisen, Chair

Evolution by natural selection is arguably the only unifying principle in biology. Despite the central importance of evolution in the study of biology, our understanding of the factors constraining the evolution of both genotypes and phenotypes is woefully incomplete. Traditionally, these questions have been addressed by studying the evolution of a particular gene or process; while this approach can be informative, it is difficult to extrapolate from such studies to more general conclusions about how life evolves. In the past several years, however, the publication of myriad genome-scale data sets has made it possible to address such questions in an unbiased, genome-wide fashion. This work represents my efforts to use these data to answer several questions in molecular evolution. I describe three factors (number of protein interactions, modularity, and coevolution) that each contribute to the evolutionary dynamics of proteins in different

ways. I also explore two phenomena, gene expression stochasticity (or "noise") and aging, and show that it is possible and informative to examine their evolution at a genomic scale. In sum, this work shows that a great deal of information can be gleaned from previously published data sets, when one examines novel questions that have not been addressed previously, and in this way can be seen as an example of this relatively new approach to biological research. It also represents an advancement in our knowledge of how life has been, and still is, evolving.

Acknowledgements

I dedicate this dissertation to the many mentors who have helped me along the way: to those who have been with me from the beginning, my parents Janis Fraser and Tom Fraser; to the people who first introduced me to research, Scott Snapper and David Sinclair; to my collaborators who have taught me a great deal, Joshua Weitz, Aaron Hirsh, and Joshua Plotkin; to the people at Berkeley who have made my time here so enjoyable and stimulating, Michael Kobor, Jasper Rine, Michael Botchan, Nicholas Cozzarelli, Jessica Shugart, Peter Dixon, and the members of the Eisen lab; and finally to my graduate advisor Michael Eisen, who was always supportive, but never overbearing.

Table of Contents

Page

Chapter I.

Introduction

Evolution

A traditional view that permeates much of science has been that mathematics is the most fundamental of scholarly disciplines, followed by physics, then chemistry, biology, and finally the "social sciences" at the proverbial bottom of the barrel. This perspective has been reinforced by the notion that each discipline can be partially or wholly explained by principles from the previous rungs of the ladder; for example, much of biology can (at least in theory) be reduced to chemistry, and chemistry can be reduced to physics.

Evolution by means of natural selection, however, does not fit well into any location on this hierarchy. Perhaps the most obvious location for it would be to lump it in with biology, since the study of evolution is usually synonymous with the comparison of different biological systems. But this would be giving short shrift to Darwin's brainchild, since evolution by natural selection is a concept far more general than anything having only to do with life.

This can be demonstrated by examining the logical prerequisites for achieving evolution by natural selection. First, there must be reproduction with heritable variation; if no offspring are produced (no reproduction), or offspring do not resemble their parents any more than random (no heritability), or all offspring are identical (no variation), then it can be seen that there is nothing for selection to act upon. If these three conditions are met, however, it follows directly that those lineages which consistently reproduce at the fastest rate will increase in numbers and eventually out-compete their disadvantaged rivals. Nothing more must be specified or added to the mix to achieve natural selection. (One hidden assumption here is that there are finite resources, since if energy is infinite,

then all lineages would continue to live [though some still more than others]; however since any system one chooses to study must have only finite resources, then this assumption is easily met and can safely be ignored). Because there is no reference to living systems in this complete description of the requirements for natural selection, then it follows that while all life evolves, all that evolves is not necessarily alive.

Several examples may help to demonstrate this point. Other instances of evolution by natural selection include entities whose study is the realm of social sciences, such as ideas: good ideas (sometimes referred to in this context as "memes"; Dawkins 1976) spread from one mind to another by means of communication, in contrast to poor (that is, less fit) ideas, which are far less likely to spread. Similarly, cultures (which can be thought of as complex aggregates of ideas and beliefs) that tend to conquer others (by whatever means, violent or otherwise) will become more prevalent, by means of natural selection; the languages these cultures bring with them evolve as well, and phylogenetic methods imported from biological taxonomics have even been applied to study language evolution. Another example of evolution by natural selection involves computer programs: software can be written that mutates and competes with other programs for a limited resource, such as computer processor time. While such programs are most often used as a means to study biological evolution, they are themselves a bona fide example of evolution (Ray 1991). Finally, it has even been suggested that if multiple universes exist (a widely accepted idea in modern physics; Smolin 1997) and are able to give rise to yet more universes that resemble their "parents" in the values of certain fundamental physical constants, then lineages of these more prolific universes would increase in frequency, having been naturally selected for their ability to propagate themselves (Smolin 1997).

While this theory of selection acting on multiple universes is currently only an unsubstantiated hypothesis, it nevertheless serves as a suitable example of how evolution by natural selection is by no means confined to biology.

As demonstrated by the reasoning and examples above, evolution by natural selection applies to many levels of the traditional hierarchy of thought. Therefore it follows that no mental shoehorn could possibly coax evolution into the confines of a single discipline; evolution transcends biology. Very few other concepts share such broad instances as to so utterly defy classification.

What is the significance of this apparent generality of the concept of evolution by natural selection? It means that one may be able to apply insights gained by studying evolution in one system, to other systems as well. For example, perhaps understanding how and why most species have gone extinct may shed some light on how and why most cultures have met a similar fate (though caution must of course be taken in not over-extending the generality of results). In order to ever hope to achieve such broad conclusions, however, evolution must be studied as systematically as possible. While understanding how a single gene's evolution has contributed to an organism's phenotype may be invaluable for understanding that particular phenotype, it is unlikely to shed any light on the general properties of evolution (in any system, biological or otherwise). What is needed to advance out understanding of the process of evolution as a whole is a means to study natural selection in a systematic, unbiased fashion; only then could the results even possibly be used to understand disparate evolving systems. For the vast majority of the time since Darwin first published his revolutionary idea (Darwin 1859), this type of systematic biological study has been simply impossible. However this has all

changed in just the past several years, with the advent of a radically new field of biology: genomics.

Genomics

Traditionally, biologists have pursued research on a particular protein, pathway, or other small sliver of an organism's biology. Innumerable great contributions to biology have been made in this fashion, and most biological research continues to be done this way. Recently, the development of high-throughput technologies has spawned a field with a very different level of focus: genomics, which is the simultaneous study of all genes in an organism. This type of research requires a fundamentally different viewpoint than traditional biology; most discoveries in this new area deal with cataloging the functions or other attributes of thousands of genes simultaneously, without achieving a deep understanding of any one gene or pathway in particular. Therefore it is pointless to compare the merits of the "old" versus the "new" biology, as both provide important but quite different contributions to biological knowledge. Many biologists have embraced this new paradigm of high-throughput biology, and as a result the past several years have witnessed an explosion in the production rate of genome-scale data (including genome sequences, genome-wide gene expression data, protein interaction maps, genome-wide synthetic lethal screens, and much more). This infusion of large data sets has, in turn, led the field of biology to a unique turning point: for the first time, a major limiting factor for advancement is analysis rather than production of data.

This need for analysis has spawned yet another new sub-discipline, known as bioinformatics, riding on the coattails of genomics. Bioinformatics concerns the analysis of the copious amounts of data constantly being produced with high-throughput genomic technologies. The two fields enjoy a close mutualistic relationship, as genomic data can

tell us little without proper analysis, and analysis cannot be done without a steady flow of data.

It is this field of data analysis to which the following chapters are devoted. More specifically, I have explored the interface between evolutionary biology and bioinformatics, in the hope of gleaning some general features of molecular evolution that could not have been discovered without a systematic (i.e., genomic) vantage point. Because the vast majority of genome-scale data sets are published by biologists without any consideration of their evolutionary implications, there are many fascinating evolutionary questions that can be addressed with these data, subsequent to their initial publication. The ubiquity of evolution as the process that has shaped the phenotypes and genotypes of every species is perhaps its greatest strength in terms of the potential contributions of diverse large-scale data sets to our understanding of this process: literally any measurement performed on an organism is shedding light, in one way or another, on that organism's evolution. By focusing on the level of genomes rather than particular genes, I have been able to extract evolutionary information from these data sets in a relatively unbiased fashion, and uncover general rules of evolution that apply more broadly than most previous discoveries in the field.

Below I describe five general categories of evolutionary genomic research that I have pursued in my graduate studies. The major theme is that each of these reflects a different way to utilize previously published data to study evolution; there is little commonality in the specific questions asked, or the results obtained (and because of this diversity of topics, I leave more detailed introductions on each area to the beginning of each chapter). As genome-scale data sets continue to accumulate at an ever-faster rate, it

is likely that more and more resources will be spent on analysis of existing data, rather than production of new data; these results illustrate just a tiny, but hopefully interesting, fraction of what can possibly be done with only the information already at our fingertips.

Chapter II

Protein-protein interactions and evolutionary constraints.

The majority of this chapter was previously published as three papers:

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. *Science* 296: 750 (2002).

Fraser HB, Wall DP, Hirsh AE. *BioMedCentral Evolutionary Biology* 3: 11 (2003).

Fraser HB, Hirsh AE. *BioMedCentral Evolutionary Biology* 4: 13 (2004).

Evolutionary rate in the protein interaction network

Abstract

High-throughput screens have begun to reveal the protein interaction network that underpins most cellular functions in the yeast *Saccharomyces cerevisiae*. How the organization of this network affects the evolution of the proteins that compose it is a fundamental question in molecular evolution. We show that the connectivity of well-conserved proteins in the network is negatively correlated with their rate of evolution. Proteins with more interactors evolve more slowly not because they are more important to the organism, but because a greater proportion of the protein is directly involved in its function. At sites important for interaction between proteins, evolutionary changes may occur largely by coevolution, in which substitutions in one protein result in selection pressure for reciprocal changes in interacting partners. We confirm one predicted outcome of this process--namely, that interacting proteins evolve at similar rates.

Introduction

A protein's rate of evolution is thought to depend both on its dispensability to the organism and on the proportion of potential amino acid changes that are compatible with proper protein function (Wilson et al 1977). We recently analyzed functional genomic data (Winzeler et al 1999) in conjunction with genomic comparisons (Chervitz et al 1998) to confirm and further characterize the relation between protein dispensability and evolutionary rate (Hirsh and Fraser 2001). Here we apply a similar approach to investigate how protein function constrains evolution. Early studies of the structure and

function of individual proteins suggested that, because molecular interactions require precisely specified structures, they impose constraints on sequence evolution (Dickerson 1971; Zuckerkandl 1976). Recent advances in the rapid detection of protein-protein interactions (Schwikowski et al 2000; Uetz et al 2000; Ito et al 2001), as well as in the sequencing of complete genomes, allow us to expand the scale on which the evolutionary effects of molecular interactions are investigated and shift from a focus on individual proteins to a broad survey of the proteome and characterization of the general relation between protein interaction and evolution.

Results and Discussion

We compiled a list of 3541 interactions between 2445 different yeast proteins (Uetz et al 2000; Ito et al 2001; Mewes et al 2002). To estimate the evolutionary rates of these proteins, we compared putatively orthologous sequences between *Saccharomyces cerevisiae* and the nematode *Caenorhabditis elegans*. A subclass of putative orthologs, which we called "well-conserved orthologs," exhibited >50% amino acid identity over aligned regions; 1531 sequence pairs met our criteria for putative orthologs, and 309 of these were in the well-conserved class. For each pair of orthologs, we estimated the evolutionary distance (K) that separates the two sequences, where K is defined as the number of substitutions per amino acid site that have taken place since the fungi-animal split. There were 164 yeast proteins for which we had both an estimate of the number of interactors and a well-conserved ortholog in the nematode. Among these proteins, there is a significant negative correlation between each protein's number of interactors I and protein evolutionary rate, as estimated by distance K [Fig. 1; linear regression: K = -

0.0175I + 0.8995, Pearson's r_{IK} = -0.24, P = 0.002; Spearman's rank correlation r_{IK} = -0.21, P = 0.007]. We have corroborated this relation between protein interaction and rate of evolution with data from two recent studies (Ho et al 2002; Gavin et al 2002) that were not considered in our initial compilation of protein interactions (not shown).

Interactions could reduce evolutionary rate in two distinct ways (Fig. 2). First, if different interactions depend on different sites, proteins with more interactors could evolve more slowly because a greater proportion of the protein is involved in protein functions (Fig. 2, arrow a). Alternatively, if proteins with many interactors have a greater effect on organism fitness, they could evolve more slowly, not because a greater proportion of the sequence is required for proper function, but because the entire sequence is subject to stronger selection against slightly deleterious mutations (Hirsh and Fraser 2001). Under this hypothesis, the correlation shown in Fig. 1 emerges because a protein's number of interactors is correlated with its effect on organism fitness, which in turn affects rate of evolution (Fig. 2, arrows b and c). To determine which of these two hypotheses provides a more likely explanation for the correlation between number of interactors and evolutionary rate, we analyzed our data on interactions and evolutionary rate in conjunction with results from genetic footprinting (Smith et al 1996) and parallel analysis (Winzeler et al 1999), high-throughput methods for estimating the growth rates of yeast strains in which a single gene has been disrupted or deleted. As expected in view of the recent demonstration that highly interactive proteins are more likely to be required for viability (Jeong et al 2001), we found that a protein's fitness effect F, estimated as the reduction in relative growth rate due to deleting or disrupting the gene that encodes the protein, is positively correlated with that protein's number of interactors I; with fitness

effects measured by parallel analysis for 2235 proteins for which interaction data were available, $r_{IF} = 0.15$, $P = 3.4 \times 10^{-13}$. In addition, among all putative orthologs, evolutionary rate is negatively correlated with fitness effect (Hirsh and Fraser 2001); with parallel analysis data for 1484 yeast proteins with putative orthologs, $r_{FK} = -0.13$, $P = 4.3 \times 10^{-7}$. Thus, among all putative orthologs, both correlations required by our second hypothesis are present: Number of interactors is correlated with fitness effect (Fig. 2, arrow b), which is correlated with evolutionary rate (Fig. 2, arrow c).

However, when we consider only well-conserved orthologs, for which the correlation between protein interaction and evolutionary rate is strongest (Fig. 1), no relation between fitness effect and evolutionary rate (Fig. 2, arrow c) is detected. Therefore, protein fitness effect is very unlikely to mediate the correlation between protein interaction and evolutionary rate. We can confirm this conclusion statistically by using parametric (Kaplan 2000) and nonparametric (Gibbons 1993) partial correlation to estimate the correlation between number of interactors and evolutionary rate while fitness effect is held constant. The parametric path coefficient ($p_{IK} = -0.25$, $P = 0.001$) and nonparametric partial measure of association (Kendall's partial $\tau_{IK} = -0.15$, $P = 0.002$) indicate a significant correlation between number of interactions and evolutionary rate that does not depend on overall protein fitness effect.

Protein sites may be involved in interactions directly, through participation in intermolecular contacts; or indirectly, through effects on overall protein conformation. In either category of sites, substitutions would be likely to perturb proper interaction and would often be removed by selection. However, removal might not occur if a substitution in one protein were followed by a complementary change in its interacting partner. In this

case, the pair of substitutions might be fixed by drift or positive selection (Rawson et al 2000; Koretke et al 2000). If such coevolution is indeed an important mode of change in proteins constrained by interactions, then interacting proteins should evolve at similar rates. We tested this prediction by examining all 411 protein interactions in which each protein had a putative ortholog in *C. elegans* and showed no significant sequence similarity with its interacting partner. For each interaction, we calculated ΔK, the difference between the evolutionary distances separating the yeast proteins from their respective orthologs in the nematode. We then averaged these differences across all 411 interactions to find the mean difference in evolutionary rate between interacting proteins, $\Delta\overline{K}^* = 1.3$ substitutions per site. To assess the significance of this difference, we repeatedly permuted our list of 411 interactions into random protein pairs and calculated the mean difference in evolutionary rate between arbitrarily paired proteins: 10,000 permutations yielded the distribution of $\Delta\overline{K}$ values shown in Fig. 3A. In all but 44 of the 10,000 permutations, our observed $\Delta\overline{K}^* < \Delta\overline{K}$, indicating that interacting proteins evolve at rates significantly closer than is expected to occur by chance ($P = 0.0044$).

Although coevolution provides an appealing explanation for the similarity in the evolutionary rates of interacting proteins, alternative hypotheses must be considered. The proteins in an interacting pair presumably act in the same functional pathway and therefore are likely to have similar effects on organism fitness. Because the dispensability of a protein influences its rate of evolution (Hirsh and Fraser 2001), the similarity in the evolutionary rates of interacting proteins could be a consequence of similarity in their fitness effects. Our test of this hypothesis involved two steps.

First, we tested whether proteins that interact do indeed have similar effects on organism fitness. A randomization test showed that the mean difference in fitness effects between interacting proteins, $\Delta\overline{F}^* = 0.41$, was significantly smaller than the mean difference between arbitrarily paired proteins $\Delta\overline{F}$ ($P < 10^{-5}$) (Fig. 3B). Thus, interacting proteins do have similar effects on organism fitness.

Second, we determined whether the observed similarity in fitness effects of interacting proteins was sufficient to explain the similarity in their rates of evolution. Path analysis based on the causal model shown in Fig. 3C indicated that the correlation between the fitness effects of interacting proteins contributes only slightly to the correlation between their evolutionary rates. Thus, similarity in fitness effects is not sufficient to explain the observed similarity in the evolutionary rates of interacting proteins.

We also considered two other alternatives to the coevolutionary hypothesis. First, interacting proteins might evolve at similar rates simply because they have similar numbers of interactors, and, as shown in Fig. 1, the number of interactors influences the rate of evolution. However, we found that proteins that interact do not have similar numbers of interactors ($r_{I1I2} = 0.02, P = 0.26$). A second possibility is that interacting proteins evolve at similar rates because they exhibit structural homology and therefore have similar distributions of constrained sites. The most likely origin of structural homology between interacting proteins is duplication of the gene that encodes a homodimeric protein, followed by evolution of one copy of the gene. This process would result in homology not only between the structures, but also between the sequences, of interacting proteins. Hence, we have ensured that none of the interactions in our data set

occur between proteins that exhibit detectable sequence similarity. Thus, to account for the similarity in evolutionary rates that we observe, structural similarity would have to be independent of sequence, which would be difficult to explain evolutionarily. In sum, having considered a number of alternative hypotheses, we conclude that the coevolution of interacting proteins may be largely responsible for the observed similarity in their rates of evolution.

Beyond describing the relation between a protein's interactions and its rate of evolution, the correlations presented here could find application in the rapid assessment of functional genomic data. Much as gene expression levels have recently been used to assess protein-protein interaction data sets (Grigoriev 2001), the correlation between protein interaction and evolutionary rate may allow one to use simple genomic sequence comparisons to statistically assess the quality of large interaction data sets. More generally, correlations between protein interaction, fitness effect, and evolutionary rate may provide a means by which multiple bioinformatic data sets can be quickly cross-referenced to assess the reliability of any single method or data set.

Figure 1.

The relation between the number of protein-protein interactions (*I*) in which a yeast protein participates and that protein's evolutionary rate, as estimated by the evolutionary distance (*K*) to the protein's well-conserved ortholog in the nematode *C. elegans*.

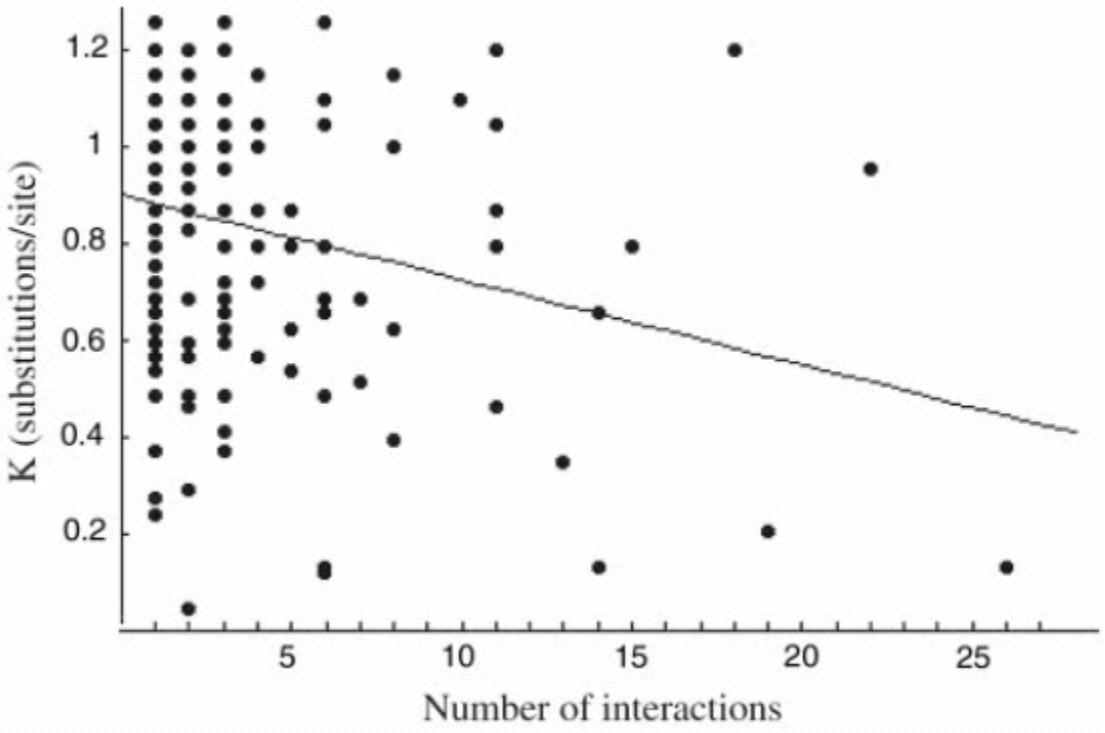

K (substitutions/site)
1.2
1
0.8
0.6
0.4
0.2
5
10
15
20
25
Number of interactions

Figure 2.

The causal model for alternative hypotheses to explain the correlation between number of interactors and evolutionary rate. One hypothesis, represented by arrow a, is that protein interactions impose structural constraints, which limit the number of substitutions that are compatible with proper protein function. A second hypothesis, represented by arrows b and c, is that proteins with more interactions have a greater effect on organism fitness and are therefore subject to stronger purifying selection. The second hypothesis can be rejected because the effect of protein interactions on evolutionary rate is not mediated by protein fitness effect.

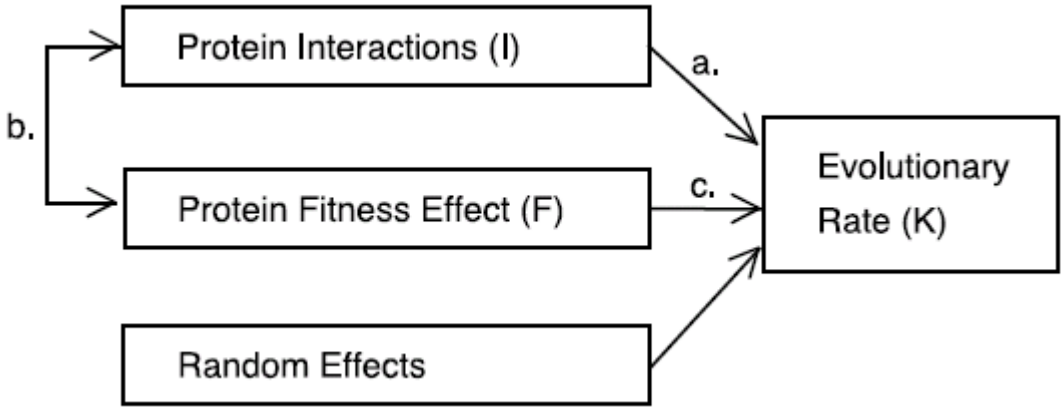
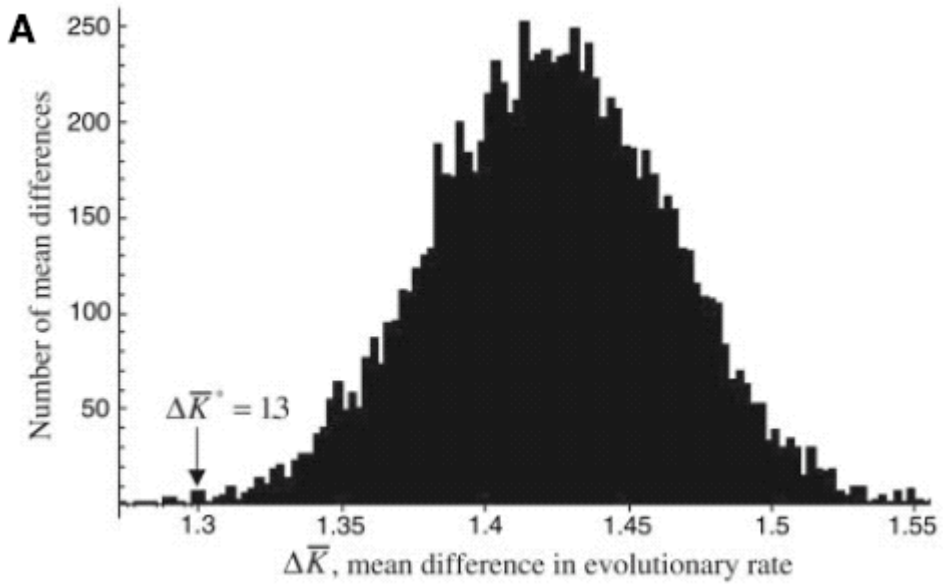
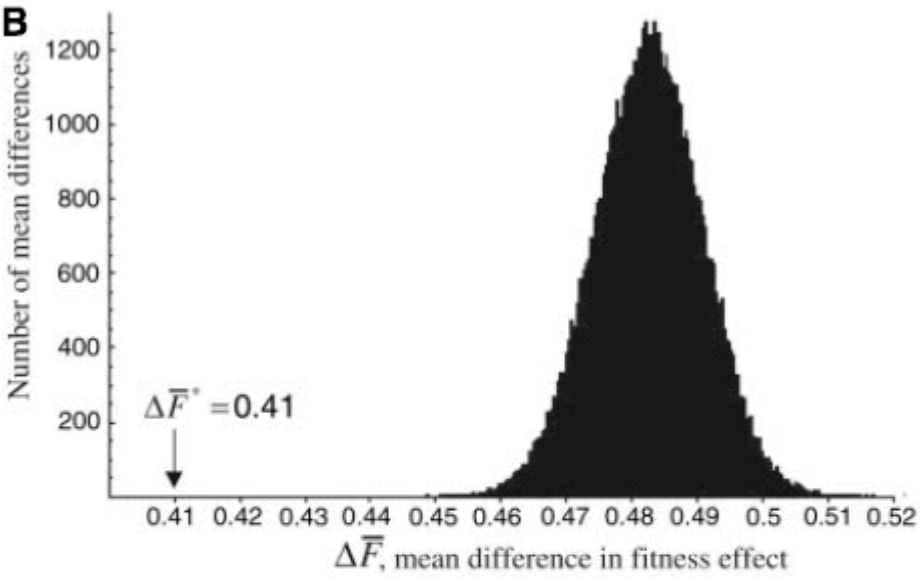
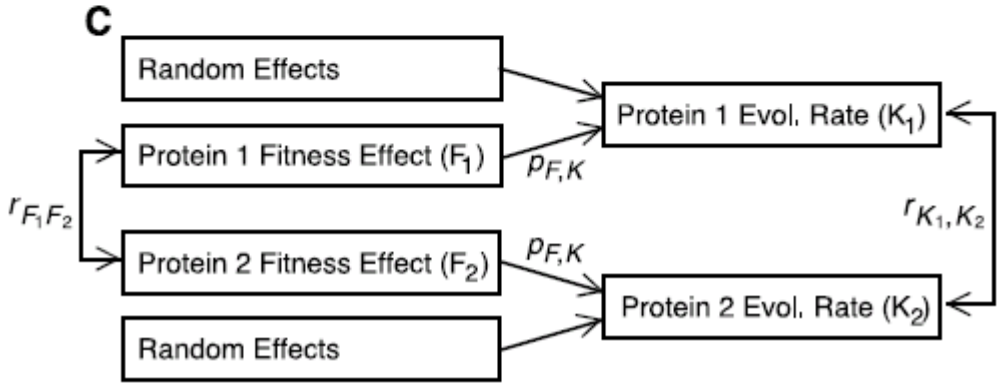

Protein Interactions (I)
a.
b.
Protein Fitness Effect (F)
c.
Evolutionary Rate (K)
Random Effects

Figure 3.

Interacting proteins have similar fitness effects, but this cannot explain the similarity in their rates of evolution. (**A**) The distribution of mean difference in evolutionary rate ($\Delta\overline{K}$) between yeast proteins randomly chosen from the list of all 411 interacting protein pairs in which both members had an ortholog in *C. elegans*. The mean difference in evolutionary rate between proteins that interact ($\Delta\overline{K}$* = 1.3 substitutions per site) is indicated by an arrow. (**B**) The distribution of mean difference in fitness effect ($\Delta\overline{F}$) between yeast proteins randomly chosen from the list of all 2821 interactions in which the effect on growth rate of deleting each protein was estimated by parallel analysis (Winzeler et al 1999). The mean difference in fitness effect between proteins that interact ($\Delta\overline{F}$* = 0.41) is indicated by an arrow. (**C**) The causal model for path analysis to determine whether similarity in fitness effects between interacting proteins explains the similarity in their evolutionary rates. The correlation between evolutionary rates of interacting proteins that is expected to result from observed correlations between fitness effects (r_{F1F2}), and between fitness effect and evolutionary rate (p_{FK}), can be estimated as $\hat{r}_{K1K2} \sim (p_{FK})^2 r_{F1F2}$. The observed correlation between evolutionary rates is much larger than that expected to result from fitness effects ($r_{K1K2} >> \hat{r}_{K1K2}$), indicating that one or more additional factors must contribute to the similarity of evolutionary rates of interacting proteins. (Observed correlation coefficients, including essential proteins: $r_{F1F2} = 0.16$, $P < 10^{-15}$; $p_{FK} = -0.06$, $P = 0.02$; $r_{K1K2} = 0.11$, $P = 0.03$. Excluding essential proteins: $r_{F1F2} = 0.07$, $P = 0.01$; $p_{FK} = -0.14$, $P = 2 \times 10^{-5}$.)


A
250
200
150
100
50
Number of mean differences
$\Delta\overline{K}^{*} = 1.3$
1.3
1.35
1.4
1.45
1.5
1.55
$\Delta\overline{K}$, mean difference in evolutionary rate


B
1200
1000
800
600
400
200
Number of mean differences
$\Delta\overline{F}^{*} = 0.41$
0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49 0.5 0.51 0.52
$\Delta\overline{F}$, mean difference in fitness effect


C
Random Effects
Protein 1 Evol. Rate (K_1)
Protein 1 Fitness Effect (F_1)
$p_{F,K}$
$r_{F_1F_2}$
r_{K_1,K_2}
Protein 2 Fitness Effect (F_2)
$p_{F,K}$
Protein 2 Evol. Rate (K_2)
Random Effects

A simple dependence between protein evolution rate and the number of protein-protein interactions

Abstract

It has been shown for an evolutionarily distant genomic comparison that the number of protein-protein interactions of proteins correlates negatively with their rates of evolution. However the generality of this observation has recently been challenged. Here we examine this problem using protein-protein interaction data from the yeast *Saccharomyces cerevisiae* and genome sequences from two other yeast species. In contrast to a previous study that used an incomplete set of protein-protein interactions, we observed a highly significant correlation between number of interactions and evolutionary distance to either *Candida albicans* or *Schizosaccharomyces pombe*. This study differs from the previous one in that it includes all known protein interactions from *S. cerevisiae*, and a larger set of protein evolutionary rates. In both evolutionary comparisons, a simple monotonic relationship was found across the entire range of the number of protein-protein interactions. In agreement with our earlier findings, this relationship cannot be explained by the fact that proteins with many interactions tend to be important to yeast. The generality of these correlations in other kingdoms of life unfortunately cannot be addressed at this time, due to the incompleteness of protein-protein interaction data from organisms other than *S. cerevisiae*. In sum, protein-protein interactions tend to slow the rate at which proteins evolve. This may be due to structural constraints that must be met to maintain interactions, but more work is needed to definitively establish the mechanism(s) behind the correlations we have observed.

Introduction

What factors determine the rates at which different proteins evolve is a fundamental question in molecular evolution. With the advent of functional genomics, this question can now be addressed on a genome-wide scale. Different determinants of evolutionary rate revealed by analysis of functional genomic data include protein dispensability (Hirsh and Fraser 2001), transcript level (Pal et al 2001), and number of protein-protein interactors (Fraser et al 2002).

Recently, Jordan *et al*. (2003) suggested that the correlation between a protein's evolutionary rate and its number of protein interactions arises only because a few, highly interactive proteins evolve more slowly than all other proteins. In our original analysis, a distant genomic comparison of *S. cerevisiae* with *C. elegans* was used to find approximate evolutionary rates of putatively orthologous genes shared by these two species. One would expect that comparisons of more closely related species would increase the strength of the relationship, since more orthologs can be found and evolutionary rates can be estimated with greater precision. Surprisingly, when Jordan *et al*. compared orthologs between *S. cerevisiae* and another yeast, *S. pombe*, they found only an extremely weak relationship between number of protein interactions and evolutionary rate. Furthermore, they found that when proteins were binned by their number of interactions, only the bin containing the most highly interactive proteins showed any reduction in evolutionary rate. They concluded from this that there is no general correlation between number of protein interactions and evolutionary rate, and that the reduction of evolutionary rate observed in the most highly connected proteins may be

an indirect effect of the relationship between protein dispensability and rate of evolution (Jordan et al 2003).

Here we show that the absence of a general correlation between protein interactions and evolutionary rate in the analysis of Jordan *et al.* can be attributed to an incomplete dataset. Our analysis differs from that of Jordan *et al.* in two basic ways. First, Jordan et al. used only protein-protein interactions from the MIPS database (Mewes et al 2002), which consists of individually reported interactions combined with data from the high-throughput screen of Uetz *et al*. (2000). While the MIPS database contains many high-confidence interactions, it is very small when compared to the total number of interactions known from all high-throughput screens. Second, Jordan *et al*. identified orthologs by taking reciprocal best BLAST hits, a method that leads to an incomplete list, because the top BLAST hit is often not the most closely related protein (Koski and olding 2001). We used a method based on maximum likelihood estimation of evolutionary distances that results in a more complete list (Wall et al 2003).

Using our more complete lists of both protein-protein interactions and orthologs, we show here that the correlation we originally reported in the *C. elegans – S. cerevisiae* comparison is indeed even stronger when more closely related genome sequences are compared. We use orthologs between *S. cerevisiae* and *S. pombe,* as well as the more closely related yeast, *C. albicans*, to probe this relationship in greater detail than we did in our previous study. We find a simple monotonic relationship between number of protein interactions and evolutionary rate, and we find that this relationship applies to proteins with few interactions, as well as to those with many.

Protein-protein interactions and evolutionary rates in yeast

We compiled a list of *S. cerevisiae* protein-protein interactions from every major high-throughput study published to date (Uetz et al 2000; Ito et al 2001; Gavin et al 2002; Ho et al 2002), as well as individually reported interactions from the MIPS database (Mewes et al 2002). The final non-redundant set consists of all interactions used in our previous study (Fraser et al 2002), and contains 13,925 interactions involving 3575 proteins. This is a more comprehensive data set than that analyzed by Jordan *et al*. (2003), which contained fewer than 2500 interactions once duplicate interactions were removed (I.K. Jordan, personal communication).

Using the genome sequences of *C. albicans* and *S. pombe* for comparison with *S. cerevisiae*, we identified putative orthologs using a maximum likelihood-based approach (Wall et al 2003), which identified 3727 orthologs between *S. cerevisiae* and *C. albicans*, and 2988 orthologs between *S. cerevisiae* and *S. pombe*. All data will be made available upon request.

Taking the intersections of our interaction and ortholog data sets, we plotted the number of protein-protein interactions vs. evolutionary rate for all genes for which we had both types of data. For all genes in the *S. pombe*-*S. cerevisiae* comparison, we found a highly significant relationship (Figure 1a; n=2119, Spearman Rank r=-0.24, P=5.8x10^{-30}). This correlation is stronger than the rank correlation that we reported in our original study (Fraser et al 2002), and is over 27 orders of magnitude more statistically significant, due to both the increased strength and the far greater number of genes involved. Thus our expectation of a more significant correlation from a closer genomic comparison is borne out by the data.

A conclusion of Jordan *et al*. (2003) was that only the proteins with the most interactors showed any reduction in evolutionary rate–i.e., the relationship between interactions and evolutionary rate was confined to those proteins with the most interactors. As shown in Figure 1b, when a more complete set of interactions and orthologs is used, the relationship can be seen to extend over the entire range of number of interactions. It takes the form of a simple monotonic relationship. This supports the idea that regardless of how many protein-protein interactions a protein participates in, each interaction affects the protein's rate of evolution.

This same analysis can be repeated using an *S. cerevisiae*-*C. albicans* genomic comparison, and the same set of *S. cerevisiae* protein-protein interactions. When we perform this analysis, the results are even stronger than for the *S. pombe* comparison. As shown in Figure 2a, a significant correlation is found (n=2496, Spearman Rank r=-0.25, P=5.2x10^{-38}). Separating the data into bins by their number of interactors also shows the same relationship as for the *S. pombe* comparison, with a clearly monotonic relationship observable over the entire range of protein interactions per protein (Figure 2b).

What is the source of the difference between the two studies?

Our finding of a strong correlation where Jordan *et al*. (2003) did not find one raises the question of what causes the difference. There are two possibilities: our lists of protein-protein interactions, or our lists of orthologs and the associated evolutionary rates. To answer this question, we first tested the correlation between our list of protein interactions and Jordan *et al*.'s list of orthologs and evolutionary rates. We observed a significant correlation between the two (Spearman Rank r=-0.22, P=8.5x10^{-24} Figure 3a),

only slightly weaker than our correlation in Figure 1a. Next we plotted Jordan *et al*.'s protein interaction data against our list of evolutionary rates. We found no significant correlation between the two data sets (Spearman Rank r=-0.01, P=0.79; Figure 3b). This demonstrates that the difference in our findings was due to the difference in our protein interaction lists, and not in our list of orthologs or evolutionary rates, and it underscores the importance of using datasets that are as complete as possible in this type of analysis.

Is it an indirect correlation?

Jordan *et al*. speculate that the reduction in evolutionary rate of the most highly connected proteins could be due to their greater likelihood of being essential for viability of the cell (Hirsh and Fraser 2001; Jeong et al 2001). However in our original analysis we showed that the effect on cell fitness when a gene is deleted cannot explain the correlation between number of protein interactions and evolutionary rate (Fraser et al 2002). In order to investigate this question for these less distant genomic comparisons, we repeated the analysis from our original study. We used Kendall's Partial Tau (Gibbons 1993), a metric of partial correlation that allows one to quantify the magnitude of a correlation between two variables when a third, potentially related variable is statistically held constant. For example, in Figure 4a, a diagram is shown in which the arrows connecting the three variables represent the relationships among them. We used Kendall's Partial Tau to assign a P-value (by 10^5 randomization tests of the data) to each arrow, representing the probability that the arrow represents a correlation that is significantly different from zero when the third variable is statistically controlled. When we use this method to analyze the number of protein-protein interactions, evolutionary

rate, and fitness effect of each gene, we find that fitness effect cannot explain the relationship between number of protein-protein interactions and evolutionary rate for either of our genomic comparisons (Figure 4b-c), consistent with our original study (Fraser et al 2002).

Protein-protein interactions and evolutionary rates in bacteria

Jordan *et al*. (2003) also note that they cannot detect a correlation between number of protein-protein interactions and evolutionary rate in *Helicobacter pylori*. Based on this observation, they conclude that the relationship between interactions and evolutionary rate does not apply to bacteria. However, substantial caution must be exercised in interpreting results that are based on a single protein interaction study (Rain et al 2001). Indeed, when using either one of the first two published high-throughput yeast protein interaction data sets (Uetz et al 2000; Ito et al 2001) alone, it is not possible to find a significant correlation between the number of interactors and evolutionary rate; it is only through a compilation of several data sets that a significant relationship emerges for yeast. Until this is possible for *H. pylori*, we should be reluctant to conclude whether or not such a relationship exists in this organism.

Conclusions

We have shown that the previously reported relationship between protein-protein interactions and evolutionary rates of proteins is even stronger when comparing different yeast species than it is when comparing yeast with *C. elegans*. The fact that the

relationship can be detected at all with a genomic comparison of species separated by approximately one billion years of evolution (*S. cerevisiae* and *C. elegans*), as well as with the comparisons of the more closely related species presented here, underscores the robustness of the relationship. That the correlation cannot be detected when using a smaller set of protein-protein interactions, as in the study by Jordan *et al*. (2003), demonstrates the importance of using data that are as complete as possible when correlating diverse genomic data. Since no such complete data set is available for any organism other than *S. cerevisiae*, it is not yet possible to judge whether the relationship applies to prokaryotes as well as eukaryotes.

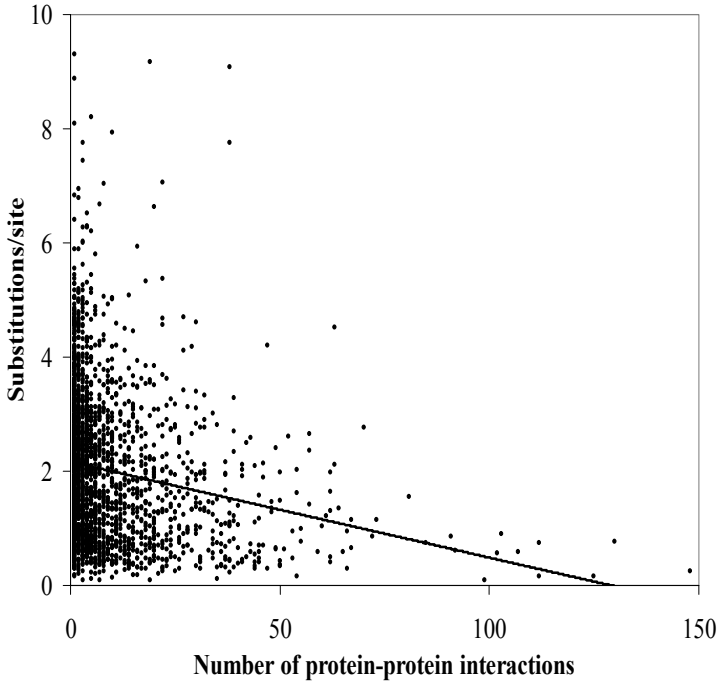
It was correctly noted by Jordan *et al*. that the correlation we previously observed explains only about 6% of the variance in evolutionary rates (Fraser et al 2003); the correlations presented here are only slightly stronger. However, when one considers the various and unavoidable sources of noise in the analysis (e.g., identifying orthologs, aligning orthologs, estimating evolutionary distances, and perhaps most importantly, false positives and negatives in the protein-protein interaction data), as well as confounding biological factors (e.g., the fact that protein-protein interactions will not be invariable between the species whose genomes are compared, so interactions recently evolved in the *S. cerevisiae* lineage will not show a significant effect on evolutionary rate), it seems surprising that the correlations are as strong as they are. In view of the sources of noise presently unavoidable in evolutionary analysis of functional genomic data, the fraction of variance in evolutionary rate that is explained by any one functional paramater—such as protein interactions, dispensability, or expression—cannot yet be taken as an accurate estimate of the relative importance of that factor's role in determining the rate of

evolution. It will be interesting to see how much the strength of the correlations examined here increases, and whether the relationships take informative functional forms, as more high-quality protein-protein interaction data sets and genome sequences are published.

Figure 1.

The relationship between number of protein-protein interactions and evolutionary rate between *S. cerevisiae* and *S. pombe*. (a) The relationship between number of protein-protein interactions and evolutionary rate for all 2119 orthologs with protein interaction data. Several outliers are not shown but were included in the analysis. (b) Average evolutionary rates of genes binned by their number of protein-protein interactions.

(A)


10
8
6
4
2
0
Substitutions/site
0
50
100
150
Number of protein-protein interactions

(B)

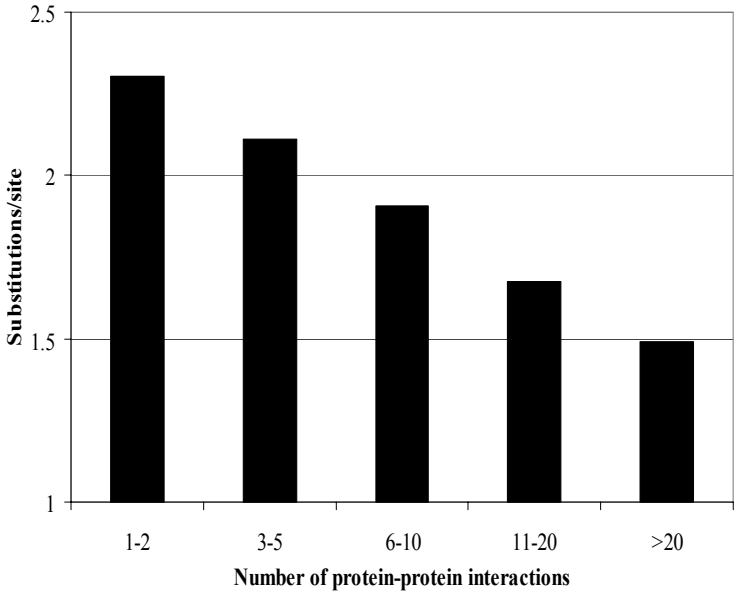
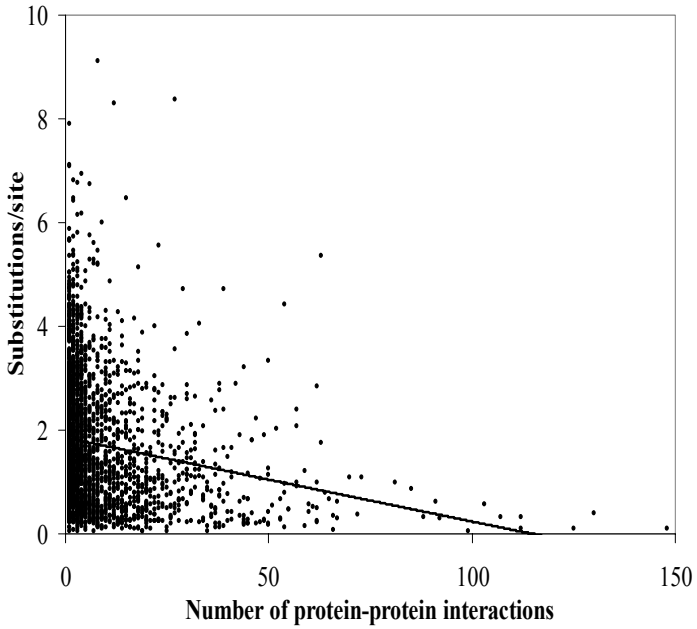

2.5
2
1.5
1
Substitutions/site
1-2
3-5
6-10
11-20
>20
Number of protein-protein interactions

Figure 2.

The relationship between number of protein-protein interactions and evolutionary rates between *S. cerevisiae* and *C. albicans*. (a) The relationship between number of protein-protein interactions and evolutionary rate for all 2496 orthologs with protein interaction data. Several outliers are not shown but were included in the analysis. (b) Average evolutionary rates of genes binned by their number of protein-protein interactions.

(A)


10
8
6
4
2
0
Substitutions/site
0
50
100
150
Number of protein-protein interactions

(B)

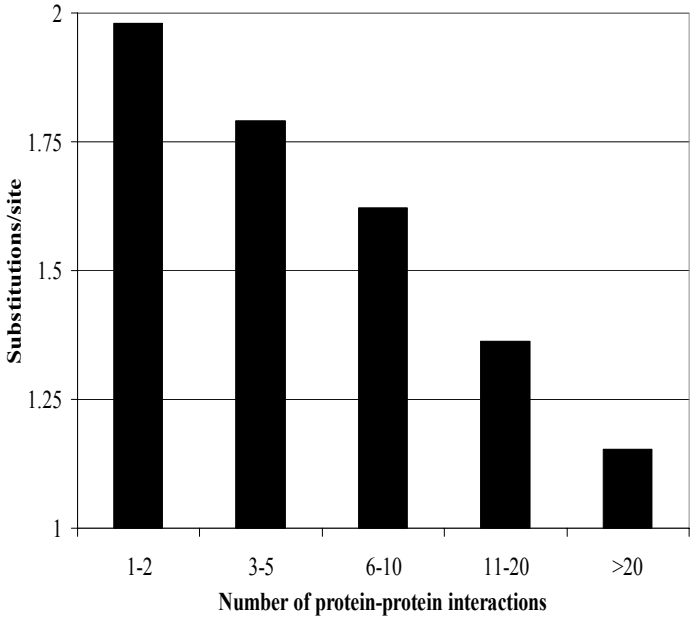
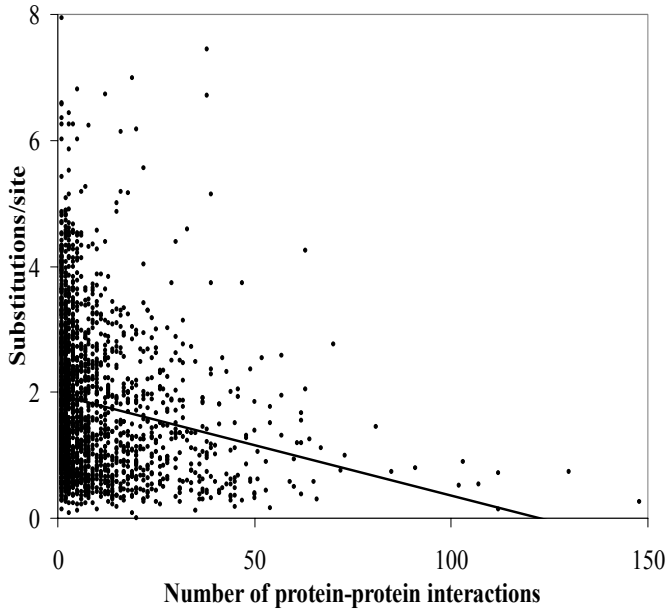

2
1.75
1.5
1.25
1
Substitutions/site
1-2
3-5
6-10
11-20
>20
Number of protein-protein interactions

Figure 3.

Testing the different lists of protein-protein interactions and evolutionary rates from the two studies. (a) A significant correlation is found when using evolutionary rates of orthologs from Jordan *et al*. (2003) with our list of protein-protein interactions. Several outliers are not shown but were included in the analysis. (b) No correlation is seen when using our evolutionary rates of orthologs with Jordan *et al*.'s list of protein-protein interactions.

(A)


Substitutions/site
0
2
4
6
8
0
50
100
150
Number of protein-protein interactions

(B)

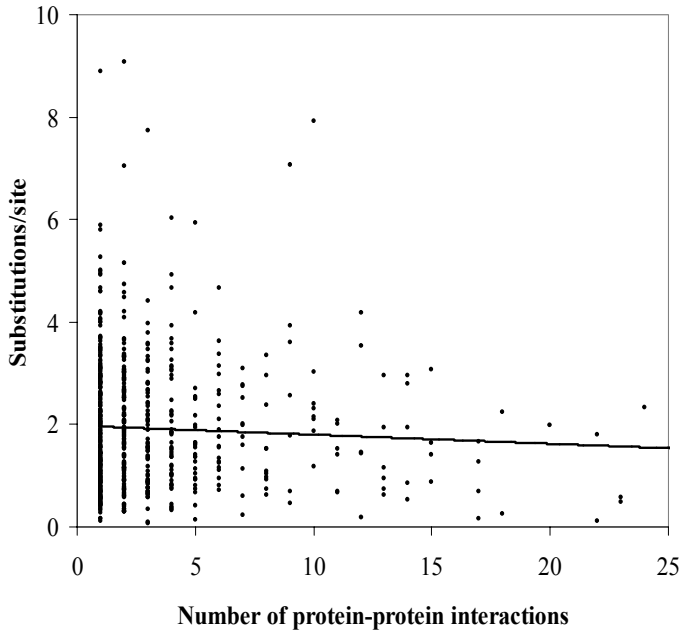
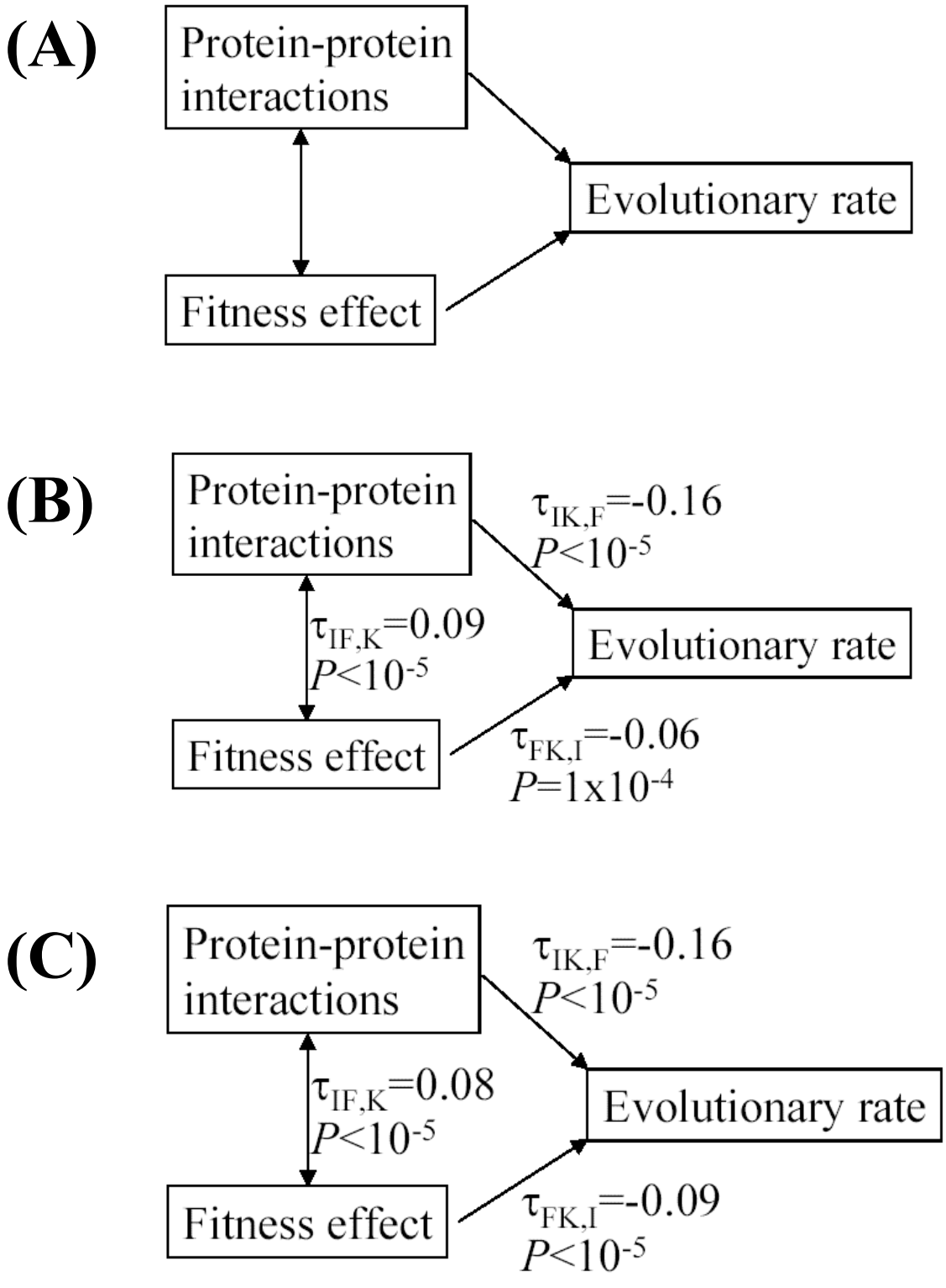

Substitutions/site
0
2
4
6
8
10
0
5
10
15
20
25
Number of protein-protein interactions

Figure 4.

Diagram of correlations between number of protein-protein interactions, evolutionary rates, and fitness effects (a) Each arrow represents the correlation between the two variables it connects. Whether or not the correlation is statistically significant by Kendall's Partial Tau is shown by the *P*-values next to each arrow in (b) and (c). (b) The significance of each correlation for the *S. cerevisiae-S. pombe* comparison. Note that the arrow connecting number of protein-protein interactions and evolutionary rates is highly significant, with none of the 10^5 randomizations of the data having a stronger correlation. (c) The significance of each correlation for the *S. cerevisiae-C. albicans* comparison. Note that the arrow connecting number of protein-protein interactions and evolutionary rates is highly significant, with none of the 10^5 randomizations of the data having a stronger correlation.


(A)
Protein-protein interactions
Evolutionary rate
Fitness effect
(B)
Protein-protein interactions
$\tau_{IK,F}$=-0.16
P<10^{-5}
$\tau_{IF,K}$=0.09
P<10^{-5}
Evolutionary rate
Fitness effect
$\tau_{FK,I}$=-0.06
P=1x10^{-4}
(C)
Protein-protein interactions
$\tau_{IK,F}$=-0.16
P<10^{-5}
$\tau_{IF,K}$=0.08
P<10^{-5}
Evolutionary rate
Fitness effect
$\tau_{FK,I}$=-0.09
P<10^{-5}

Evolutionary rate depends on number of protein-protein interactions independently of gene expression level

Abstract

Whether or not a protein's number of physical interactions with other proteins plays a role in determining its rate of evolution has been a contentious issue. A recent analysis suggested that the observed correlation between number of interactions and evolutionary rate may be due to experimental biases in high-throughput protein interaction data sets. The number of interactions per protein, as measured by some protein interaction data sets, shows no correlation with evolutionary rate. Other data sets, however, do reveal a relationship. Furthermore, even when experimental biases of these data sets are taken into account, a real correlation between number of interactions and evolutionary rate appears to exist. A strong and significant correlation between a protein's number of interactions and evolutionary rate is apparent for interaction data from some studies. The extremely low agreement between different protein interaction data sets indicates that interaction data are still of low coverage and/or quality. These limitations may explain why some data sets reveal no correlation with evolutionary rates.

Background

Over twenty-five years ago, a number of authors suggested that a protein's rate of evolution should decrease with the number of molecular interactions in which it participates (Dickerson 1971; Ingram 1961; Wilson et al 1977). The rationale behind this prediction was that additional interactions impose functional constraints on otherwise

relatively unconstrained residues, such as those on the surface of the protein. Thus, other things being equal, a protein with more interactions would evolve more slowly. This prediction was recently corroborated by us, in the form of a negative correlation between a protein's rate of evolution and the number of other proteins with which it interacts (Fraser et al 2002). While other authors have questioned the existence of this relationship (Jordan et al 2003), we later showed that in their analysis, the absence of a correlation was due to the particular protein interaction data that they used; when all data sets available at that time were used, a very strong and statistically significant correlation was apparent (Fraser et al 2003).

In a recent, thorough analysis of protein interaction data sets, Bloom and Adami have questioned whether the correlation between number of protein interactions and evolutionary rate is independent of gene expression level (Bloom and Adami 2003). While we agree that the results of Bloom and Adami show quite convincingly that an association between expression and number of interactions contributes significantly to the correlation between interactions and evolutionary rate, we believe that two of their conclusions are unwarranted. First, it is not yet clear that the association between expression and number of protein interactions is due exclusively to experimental biases rather than real properties of the organism. Second, current results do not indicate that the correlation between interactions and evolutionary rate is entirely due to the association between expression and evolutionary rate. In this work, we argue that their conclusions represent an over-extension of their analyses, and also provide further analyses demonstrating that a protein's number of interactions does indeed influence its rate of evolution, independently of its expression level.

Critique of Bloom and Adami

Bloom and Adami (2003) tested protein interaction data from seven methods (two experimental and five computational) individually for correlations between the number of protein interactions and protein evolutionary rates, while statistically controlling for gene expression levels. They found that only in the two interaction data sets generated using mass spectrometry was there a strongly significant correlation between the number of protein interactions and evolutionary rate independent of expression levels. In protein interaction data sets generated by the computational methods of gene co-occurrence and gene neighborhood, a weakly significant correlation between number of interactions and evolutionary rate remained when expression levels were statistically controlled (Bloom and Adami 2003). Despite the inability of expression levels to account for the correlation between number of interactions and evolutionary rate in these data sets, Bloom and Adami argued that expression levels completely explain the correlation between number of interactions and evolutionary rate, and that they failed to see this in the partial correlations because the partial correlations did not completely control for expression levels. To explain why partial correlations were unable to completely control for expression levels, Bloom and Adami suggested that their expression data (measured by DNA microarrays and codon bias) are imprecise.

While we agree with Bloom and Adami that current codon usage and expression data do not measure expression levels with perfect precision, we do not believe that their interpretation is supported by the evidence. If one is to consider the quality of each of the types of data involved in calculation of the partial correlations—expression data, evolutionary rate data, and interaction data—there is no question that the least reliable of

the three are the interaction data. This can be seen in many ways, the simplest of which is the nearly nonexistent overlap between different high-throughput protein interaction data sets (von Mering et al 2002). Regardless of whether this small overlap is predominantly due to false positives, false negatives, or simply incomplete coverage, the fact is that the two independent expression data sets used by Bloom and Adami show much better agreement than any two high-throughput interaction data sets in existence. (As reported by Bloom and Adami [2003], their two expression data sets are correlated with Spearman rank $r = 0.62$; in contrast, the correlation between number of interactions per protein in two of the most comprehensive and highest quality high-throughput interaction data sets [Ho et al 2002; Gavin et al 2002] is only 0.12, and correlations between most other protein interaction data sets are weaker or even negative [Hoffman and Valencia 2003]). Expression data, we may conclude, are of significantly higher quality and/or coverage than currently available interaction data. Therefore, if one is to invoke poor data as an explanation for not observing some particular outcome of the analysis, then it should be invoked to explain why the correlations involving protein interactions are not any stronger than they presently are. More generally, if the precedent set by Bloom and Adami were to be followed, then any variable A that only partially weakens a correlation between two other variables B and C when it is statistically controlled for could be claimed to be completely responsible for the correlation between B and C, if the values of A are not known with perfect precision. While it is certainly always possible that A will completely account for the correlation between B and C when it is known with more precision, this remains speculative in the absence of any supporting evidence.

As further evidence that the correlation between number of interactions and evolutionary rate is mediated by expression level, Bloom and Adami (2003) showed that only in the interaction data sets in which the proteins with many interactions are highly expressed is there a significant negative correlation between number of interactions and evolutionary rate. Working under the assumption that the observed relationship between number of interactions and level of expression is an experimental artifact, Bloom and Adami suggested that the correlation between number of interactions and evolutionary rate is due to an experimental bias toward the detection of many interactions for highly expressed proteins. However, a simple alternative explanation must also be considered: it is entirely possible that highly expressed genes do tend to have more protein interactions than weakly expressed genes. Indeed, in addition to being found in yeast, a positive correlation between expression level and number of interactions has been reported in other organisms as well, using protein interaction detection methods (such as yeast 2-hybrid) which Bloom and Adami believe are unbiased with respect to expression levels (Cutter et al 2003). If more highly expressed proteins do tend to participate in more protein interactions, one would expect to observe precisely the pattern of correlation coefficients Bloom and Adami report. Specifically, interaction datasets of sufficiently high coverage and accuracy would reveal the (real) relationship between expression and number of interactions, as well as the relationship between evolutionary rate and number of interactions. In contrast, less accurate or complete datasets would show no such relationships. As evidence against this idea, Bloom and Adami state that Jordan *et al.* (2003) "observed no significant correlation between evolutionary rate and the number of interactions when they used a set of manually curated interactions that might be expected

to be of higher accuracy than those from any single high–throughput method." While it is true that Jordan *et al*. did not observe a significant correlation, it is not true that they relied on a set of manually curated interactions. As we previously pointed out (Fraser et al 2003), approximately half of the interactions in the list used by Jordan *et al*. (after duplicate interactions were removed) were from the high-throughput yeast 2-hybrid screen of Uetz *et al*. [2000], which has been shown to be one of the least reliable high-throughput protein interaction data sets in existence (von Mering et al 2002).

Finally, Bloom and Adami criticized the biophysical explanation we proposed (Fraser et al 2002) to explain why proteins with many interactions would tend to evolve slowly. They stated that "there is no obvious reason why residues involved in intermolecular contacts should be more evolutionary [sic] constrained than other residues with the same number of intramolecular contacts" (Bloom and Adami 2003). While this is true, it is not directly relevant to our original proposal, which was that "proteins with more interactions could evolve more slowly because a greater proportion of the protein is involved in protein functions" (Fraser et al 2002). Our proposal was not that intermolecular contacts impose more stringent constraints than intramolecular contacts, but rather that additional interactions could impose constraints on sites that are otherwise relatively unconstrained, such as residues on the surface of a polypeptide. Thus the critique presented by Bloom and Adami has no bearing on the hypothesis we proposed.

Additional analysis of the data

A simple statistical method for examining the relationship between two variables (e.g., number of interactions, I and rate of evolution, E), while partially controlling for a

third, potentially related variable (e.g., gene expression, A), is to divide the dataset into quantiles according to the controlled variable. This reduces the variance of the controlled variable relative to the other variables within each quantile, resulting in partial statistical control. This approach is complementary to partial correlation in that it the two methods can be combined, and division of the dataset into bins allows one to investigate the consistency or variation of relationships across quantiles. To emphasize that current data do not indicate that the relationship between evolutionary rate and number of interactions in mass spectrometry data is entirely mediated by expression levels, we present here a simple binning and partial correlation analysis of mass spectrometry (Ho et al 2002), expression, and evolutionary rate data. It bears restating here that correlations among separate datasets indicate that interaction data are far less accurate than expression data; therefore, noise and other limitations of data should be expected to reduce the estimated strength of the relationship between number of interactions and evolutionary rate more than they reduce the strength of the relationship between expression levels and evolutionary rate.

As Bloom and Adami (2003) noted, the proteins that are chosen to be tagged and overexpressed in mass spectrometry studies are subject to an ascertainment bias. For this reason, we used only the untagged data. We used the expression data of Wang *et al*. (2002), which was produced from more replicates than other available expression datasets and, unlike the data used by Bloom and Adami (Holstege et al 1998), was not accidentally measured in an aneuploid strain of yeast (Hughes et al 2000). For evolutionary rate data we used dN/dS values calculated from four species of the

Saccharomyces genus, with a correction for the effect of codon bias on dS (Hirsh et al 2005). We used Spearman's rank correlation for all analyses.

Before dividing the dataset into quantiles according to expression level, we measured the strength of the correlation between number of interactions and evolutionary rate for all 555 genes for which we had interaction, evolutionary rate, and expression data. The correlation between number of interactions (I) and evolutionary rate (E) was quite strong, even when controlling for expression ($r_{EI} = –0.403$, $p = 5x10^{-23}$; $r_{EI.A} = –0.277$, $p = 3x10^{-11}$; Table 1, row 1, column 1). We then partitioned the dataset into quantiles according to expression levels and calculated r_{EI} and $r_{EI.A}$ within each bin. We present results using two, three, four, and five bins. In every bin, the correlation between number of interactions and evolutionary rate is significant, even after controlling for expression levels. Perhaps even more importantly, controlling for expression levels actually *strengthens* the correlation between number of interactions and expression level in three of the bins (Table 1, underlined). In one of these bins, controlling for expression levels results in more than a two-fold improvement in the *p*-value of the correlation. In order for inaccurate expression data to explain this result, the expression data in those three bins would not only have to be noisy—they would have to be negatively correlated with the true expression levels of those genes. Since this is quite unlikely to be the case, we believe the most parsimonious explanation is that the number of interaction partners a protein has is correlated with its evolutionary rate independently of its expression level.

Summary

We agree with Bloom and Adami (2003) that the quality of high-throughput protein interaction data sets is quite variable, and that some show a correlation with evolutionary rates while others do not. However we do not believe that expression levels can account for this correlation in all data sets. To support this position, we showed that limitations of the data are likely to weaken the apparent effect of number of interactions more than they weaken the apparent effect of expression. Therefore, Bloom and Adami's suggestion that the significant contribution of expression to the relationship between number of interactions and evolutionary rate should be interpreted to mean that expression is entirely responsible for this relationship seems unwarranted. To emphasize that the measurable effect of number of interactions on evolutionary rate remains highly significant even when controlling for expression, we presented a re-analysis of mass spectrometry interaction data. Across quantiles of expression, the relationship between number of interactions and evolutionary rate, controlling for expression levels, was significant. In several quantiles, controlling for expression actually strengthened the relationship between number of interactions and evolutionary rate.

Bloom and Adami's thorough analysis shows, above all, that large-scale data sets remain woefully noisy and incomplete. While it remains possible that expression levels will ultimately account for the correlation between number of interactions and evolutionary rate once more accurate expression data are published, we find it far more likely that the vast majority of improvement will be in protein interaction data. In any

case, it will be interesting to see what relationships emerge as more (and higher quality) functional genomic data are produced.

Table 1.

Quantile analysis of mass spectrometry protein interaction data. Genes were separated into 1-5 bins based on their expression levels (Wang et al 2002). Each column is an analysis of the data set with a different number of bins. Each row is the rank of each bin's average expression level, where low rank indicates low expression. The upper number in each cell is the Spearman rank correlation coefficient between number of interactions and evolutionary rate (r_{EI}). The lower number in each cell is the partial correlation coefficient between number of interactions and evolutionary rate, controlling for expression level ($r_{EI.A}$); the three cases in which this number has a greater absolute value than r_{EI} are underlined. *, $p<0.05$; **, $p<0.005$; ***, $p<0.0005$.

	1	2	3	4	5
1	–0.403*** –0.277***	–0.223*** –0.216***	–0.182* –0.182*	–0.179* –0.179*	–0.206* –0.205*
2		–0.341*** –0.285***	–0.323*** –0.328***	–0.245** –0.272**	–0.192* –0.197*
3			–0.263*** –0.258**	–0.366*** –0.358***	–0.352*** –0.336**
4				–0.225* –0.222*	–0.357*** –0.325**
5					–0.253* –0.207*

Chapter III

Coevolution of protein sequence and expression.

The majority of this chapter was previously published as: Fraser HB, Hirsh AE, Wall DP, Eisen MB. *The Proceedings of the National Academy of Science USA,* 101: 9033 (2004)

Abstract

Physically interacting proteins or parts of proteins are expected to evolve in a coordinated manner that preserves proper interactions. Such coevolution at the amino acid sequence level is well documented, and has been used to predict interacting proteins, domains, and amino acids. Interacting proteins are also often precisely coexpressed with one another, presumably to maintain proper stoichiometry among interacting components. Here we show that the expression levels of physically interacting proteins coevolve. We estimate average expression levels of genes from four closely related fungi of the genus Saccharomyces using the codon adaptation index, and show that expression levels of interacting proteins exhibit coordinated changes in these different species. We find that this coevolution of expression is a more powerful predictor of physical interaction than is coevolution of amino acid sequence. These results demonstrate coevolution of gene expression for the first time, adding a new dimension to the study of the coevolution of interacting proteins, and underscoring the importance of maintaining coexpression of interacting proteins over evolutionary time. Our results also suggest that expression coevolution can be used for computational prediction of protein-protein interactions.

Introduction

Coevolution is an evolutionary process in which a heritable change in one entity establishes selective pressure for a change in another entity, where the entities can range from nucleotides to amino acids to proteins to entire organisms, and perhaps even ecosystems. A relatively simple and well-studied example of coevolution involves physically interacting proteins, where precise, complementary structural conformations of interacting partners are usually needed to maintain a functional interaction. If the conformation of one protein is interrupted by mutation, a compensatory change may be selected for in its interacting partner. When such compensatory changes occur, the two proteins are said to coevolve.

Coevolution of interacting amino acids and proteins has been studied intensively for more than a decade (Altschuh et al 1987; Moyle et al 1994; Pazos et al 1997; Goh et al 2000; Ramani and Marcotte 2003; Pazos and Valencia 2001; ibid 2002; Goh and Cohen 2002). The identification of coevolving pairs of genes is interesting and important for several reasons. First, it can aid in functional annotations: when an uncharacterized gene is found to coevolve with several different genes, all of which encode proteins of a single function, the unknown gene is likely to share that same function. Second, identification of likely physical interactions through detection of coevolution can contribute to our understanding of how proteins work together to execute their functions. Third, coevolution may be a critical process by which complex cellular components, such as multi-molecule machines and metabolic pathways, undergo adaptive or constructive change without disruption of organismal integrity.

Many different methods have been developed to detect coevolution of proteins, most based on a common principle: evolutionary distances between all possible pairs of amino acid sites or proteins are estimated from multiple alignments of protein sequences, and the extent of coevolution for each pair is determined by measuring the correlation of their evolutionary rates across different lineages. Such methods have been successful in quantifying the extent of coevolution between proteins, protein domains, and amino acid residues known to interact physically (Pazos et al 1997; Goh et al 2000; Ramani and Marcotte 2003; Pazos and Valencia 2001; ibid 2002; Goh and Cohen 2002). They have also been used to predict specific interactions between receptors and their substrates in large paralogous protein families (Goh et al 2000; Goh and Cohen 2002) and between proteins from the bacterium Escherichia coli (Pazos and Valencia 2001; ibid 2002).

In all previous applications of this approach to the study of protein coevolution, at least 11 sequences (and sometimes many more) have been used in each multiple alignment (Pazos et al 1997; Goh et al 2000; Ramani and Marcotte 2003; Pazos and Valencia 2001; ibid 2002; Goh and Cohen 2002). While such extensive taxonomic sampling is possible in studies of prokaryotes, for which over 100 genome sequences are currently available, it remains difficult in studies of eukaryotes.

Here we examine whether coevolution can be detected not only in protein sequences, but also in their levels of expression. The expectation that expression levels should coevolve stems in part from the observation that the expression levels of genes encoding interacting proteins are strongly correlated over different experimental conditions in *S. cerevisiae* (Eisen et al 1998; Grigoriev 2001; Ge et al 2001). This is thought to reflect the requirement for interacting proteins to be present in the cell in

similar amounts at the same time in order to properly form stoichiometric complexes and execute their function. When protein complex subunits are misexpressed, they tend to have more severe consequences on growth than proteins that do not participate in stable protein interactions (Papp et al 2003). Thus, we predicted that natural selection would maintain precise coexpression of interacting proteins; if the expression of one gene changes, it would be expected to result in a selection pressure for a similar expression change in its interacting partners, analogous to the coevolution of amino acid sequence described above.

In this study we use the genome sequences of four closely related yeasts – *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* – along with protein interaction data from *S. cerevisiae* to introduce a new method to detect coevolution of gene expression based on coordinated changes in gene expression, as estimated by codon usage bias. We also examine protein sequence coevolution, in order to evaluate whether sequence data from these four species alone allows the coevolution of interacting proteins to be detected on a genomic scale, and to compare the strength of expression coevolution to the strength of sequence coevolution.

Results

Coevolution of protein sequences

We began by examining metrics of coevolution for proteins that have been observed to interact in *S. cerevisiae*. From a set of 4175 relatively high-confidence protein-protein interactions involving 1360 proteins (von Mering et al 2002), we identified 1377 interacting pairs involving 621 proteins where both proteins had clear

orthologs in all four *Saccharomyces* species and the alignments of the protein sequences were of high quality. We used the multiple alignments to estimate rates of evolution for each protein in each lineage. For all pairs of proteins we computed - as a measure of their coevolution - the correlation coefficient between their rates of evolution in the different lineages (see Methods). For comparison to the set of interacting proteins, we generated a list of all 192,510 possible pairs (involving the same 621 proteins) that were not in our list of 1377 interactions.

Because there was a wide range in the amount of variance in evolutionary rates for different pairs of proteins (Fig 1a), we reasoned that pairs where one or both of the proteins had very little variance in evolutionary rates would not be very informative for detecting coevolution, since the small changes that are indicated by a small variance are more likely to reflect random fluctuations or noise instead of authentic changes in the evolutionary rates of a gene along different lineages. For this reason, we restricted our analysis to the 200 interacting pairs (of the 1377 total) with the greatest variance in both proteins of the pair (i.e., only the variance in the less variable of the two proteins was used to represent the pair). This variance cutoff (Fig 1a, dashed line) was then applied to the complete list of 192,510 random pairs, resulting in a list of 26,796 pairs (200 known interactions and 26,596 others) with a variance in evolutionary rates above the cutoff for every protein in the list. In other words, a minimum variance cutoff was applied to all 621 proteins, and all possible pairs among those satisfying the cutoff were included for further analysis.

If the amino acid sequences of our 200 interacting proteins were in fact coevolving, we would expect to see the distribution of correlation coefficients (our metric

of coevolution) to be greater in the 200 interacting pairs than in the 26,596 non-interacting pairs. To test this, we separated the interacting and non-interacting pairs into 10 bins each, separating protein pairs by the strength of the correlation between their sets of evolutionary rates. This analysis confirmed that we could observe such coevolution at a genomic scale: for all bins of correlation coefficients greater than or equal to the $0.4<r\leq0.5$ bin, there was a greater fraction of interacting protein pairs than random pairs (Fig 1b). These two distributions are significantly different from one another, as measured by the Kolmogorov-Smirnov (KS) test (p=0.0069). This difference can also be summarized by comparing the medians of these two distributions; as expected from Fig 1b, the median correlation coefficient for interacting pairs (r=0.088) was higher than that of random pairs (r=-0.050).

While these results establish that we can detect coevolution of interacting protein sequences using just four genome sequences, they do not quantify for what fraction of our interacting proteins we have detected coevolution. Another way to pose this same question is to ask, for what fraction of our interacting proteins do we find a correlation coefficient higher than that expected for protein pairs that are not known to interact? Since the distribution of correlation coefficients among non-interacting pairs (Fig 1b, dashed line) represents what is expected by chance, the difference between the values that form this curve and those that form our distribution of interaction correlation coefficients (Fig 1b, solid line) at high correlation coefficients (specifically, at all correlation coefficient bins greater than the largest correlation coefficient at which the distributions cross) is the value we seek. In other words, we are simply subtracting an estimate of the fraction of false positives from the fraction of true positives to find the number of true

positives not due to random chance. We calculated this value to be 0.113, indicating that we detected coevolution in the sequences of ~23 (11.3%) of our 200 interacting pairs. Since this calculation assumes that our list of interactions is free of false positives and that our non-interactor list is free of false negatives, it should be interpreted as a lower bound for the amount of sequence coevolution we can detect with four genome sequences.

Coevolution of gene expression

While our finding coevolution for 11.3% of the interacting pairs is significant, it still represents only a small fraction of the total number of interactions in our list. Thus we wished to develop a method to extract more information about protein interactions than we could from the coevolution of protein sequence alone. Since it has been shown that genes coding for physically interacting proteins tend to be coexpressed (Eisen et al 1998; Grigoriev 2001; Ge et al 2001), we reasoned that interacting proteins might show detectable coevolution of expression levels, if such coexpression must be maintained even as expression patterns change over evolutionary time.

One method to test whether expression levels coevolve would be to use DNA microarrays to measure the expression levels in a variety of species and conditions, and then to search for cases in which expression patterns of mRNAs encoding a protein and its interacting partner have changed in a coordinated fashion. Although such experiments are feasible, they are labor intensive and expensive, and we can expect the generation of expression data to lag behind genome sequencing for some time. Therefore, we asked instead if we could detect coevolution of gene expression using sequence alone. Although

we currently have no method to accurately infer patterns of expression from sequence, there does exist a very well characterized method to estimate a gene's average expression level from its sequence. Bias in the usage of synonymous codons, which was first noted over 20 years ago (Ikemura 1982), is a remarkably good predictor of average expression level. The strong association between codon bias and expression is thought to be due to selection for translational efficiency and accuracy of highly expressed genes (Akashi 2003). (Because the changes in gene expression levels we are interested in are those that occurred over the last several million years of evolution in our four *Saccharomyces* species, codon bias may reflect aspects of previous selection on gene expression that may not be apparent in microarray expression data, since microarray data are measured in laboratory conditions that are undoubtedly quite different than those of a natural yeast habitat; also for this reason, the strength of the correlation between codon bias values and microarray expression data from the laboratory cannot be taken as a precise indicator of how well codon bias predicts historical expression levels.) Since codon bias can be easily calculated for any gene sequence, we tested the hypothesis that genes encoding interacting proteins tend to coevolve in expression, and thus show coordinated changes in codon bias in different species. In other words, if codon bias for gene X is greater in species A than in species B, then we might expect codon bias for some or all genes whose protein products interact with the protein encoded by X to be greater in species A than in species B as well.

To test this hypothesis, we again began with our list of 1377 interactions among 621 proteins. We used the codon usage from the 20 most highly expressed genes in *S. cerevisiae* (Arava et al. 2003) to parameterize the codon adaptation index (CAI; see

Methods) for each species, and used the CAI to estimate expression levels for each of the 621 genes in all four species. There was a wide range of variances in CAI for the 192,510 pairs (Fig 2a), so for the same reasons described above, we restricted our attention to the 200 interacting pairs with the highest variance in CAI for both members. Application of this cutoff (Fig 2a, dashed line) to the list of all possible pairs yielded 11,781 pairs (of which 200 were known interactions and 11,581 were not).

Comparison of the distribution of correlation coefficients for the 200 interacting pairs with the 11,581 non-interactors revealed a striking difference, with the interacting pair distribution sharply skewed towards high values (Fig 2b, solid line). The median correlation coefficient for interacting pairs was 0.822, while that of non-interactors was only 0.1997. The KS test confirmed that the difference between the two distributions was quite significant ($p<10^{-26}$). Calculating the fraction of interacting pairs for which we could detect expression coevolution (as described above for protein sequence coevolution) resulted in a value of 37.3%, or ~75 of our 200 interacting pairs, which again should be interpreted only as a lower bound. Thus we were able to detect expression coevolution at a level above the random background for over a third of the interacting protein pairs.

While our finding of strong correlations between expression levels of interacting proteins in different organisms is consistent with our hypothesis of coevolution occurring by sequential mutations, another possibility must also be considered. If the genes encoding interacting proteins are often regulated by the same trans-acting factor, then a single change affecting that factor could lead to up- or down-regulation of both interacting proteins in one species. Even though this scenario does lead to correlated

changes in expression, it would not truly be coevolution. To distinguish between the true coevolution possibility and the single trans-acting mutation possibility, we utilized experimental genome-wide transcription factor binding data that are available for 113 transcription factors in yeast (Lee et al 2002). We reasoned that if single mutations in transcription factors account for some or all of our apparent expression coevolution, then genes encoding pairs of interacting proteins that are regulated by the same transcription factor should show stronger coevolution, on average, than those that are regulated by different transcription factors. Among our 1377 interacting pairs, we found 59 that were coregulated (both genes being bound by one transcription factor with a confidence of $p<0.001$). These 59 had a median CAI correlation coefficient of 0.111, significantly *less* than that of the rest of the interacting pairs (KS test $p=0.047$). While we expect that we have missed many interacting pairs that are in fact regulated by the same transcription factor (due to both false negatives in the binding data, and our not having binding data for all transcription factors), this should only serve to weaken any bias we find. Our finding that interacting pairs regulated by the same transcription factor actually have weaker coevolution than others supports our interpretation of the correlations as evidence of coevolution by sequential mutations; however we note that this analysis does not address whether those sequential mutations occurred in cis or trans. We do not have an explanation for why interacting proteins whose genes are regulated by the same transcription factor show less expression coevolution than other interacting proteins.

Prediction of protein interactions

Considering that we have two metrics that are both indicative of physical interaction between proteins, we asked if protein pairs with coevolving expression levels were more likely to show detectable protein sequence coevolution, or if instead the two metrics are largely independent. We found the latter to be the case, as the correlation between our two metrics of coevolution was extremely weak (Pearson r=0.016). Since the metrics are independent, it is possible that they could be combined to yield more information than either in isolation.

To test the power of combining the two metrics, we generated predictions of novel protein interactions. We started with the list of random protein pairs that satisfied the variance cutoffs used above for both evolutionary rates and CAI (1711 total pairs), and applied cutoffs for both correlation coefficients. We began with the arbitrary cutoffs of r>0.75 for protein sequence coevolution and r>0.9 for CAI coevolution, which yielded a list of 21 predictions (Table 1) involving proteins of both high and low CAI (ranging from 0.197 to 0.85 in *S. cerevisiae*). Of these 21 pairs, four were interactions from our list of 1377, which is 27-fold higher than expected by chance and is thus unlikely to occur randomly (p=3x10^{-5}). This enrichment can be interpreted as the approximate enrichment for interacting proteins for all pairs in the list that are not currently known to interact; in other words, each pair in Table 1 (aside from known interactors) is ~27-fold more likely to interact than a random pair of yeast proteins. More or less stringent cutoffs can also be used, to generate either more predictions with less confidence or fewer predictions with greater confidence; for example, use of a more stringent cutoff (evolutionary rate r>0.9, CAI r>0.95) on these same 1711 pairs resulted in a list of ten predictions (Table 1, first

ten rows), of which three were from our list of known interactions (42-fold enrichment, p=4x10^{-5}). These enrichments are stronger than those resulting from the application of either metric alone (data not shown), confirming our expectation that combining the two increases their power. While we could undoubtedly have improved these enrichments for known interacting pairs by testing many different cutoffs to finely tune them, one must be careful not to over-fit the data or to perform multiple tests without the appropriate statistical corrections; thus we have chosen not to do this.

It should be noted that several genes appear multiple times in the list of our predictions (Table 1), indicating that our method may prove useful at predicting small networks of interacting proteins. For example, our method predicted a fully-connected network of four proteins (Nog1p, Rlp24p, Fur1p, and Nop7p), with all six interactions of that network among our top ten predictions. Two of these interactions, namely Nog1p with Nop7p and Rlp24p, were previously known. Other predictions in this group, for example the interaction between Nop7p and Rlp24p, are quite plausible considering that they both interact with Nog1p and that such clustering of interactions within small groups of proteins is common (Goldberg and Roth 2003). Other proteins are also predicted to interact with at least one member of this group; for instance, Utp6p is predicted to interact with Nop7p, which is quite reasonable since both of these proteins are located in the nucleolus (Adams et al 2002; Dragon et al 2002). While alternative methods for computational prediction of protein interactions and functional linkages have yielded more predictions than our method, we note that they have all used far more genome sequences as well (e.g., 57 genomes were used by Date and Marcotte [2003]). Thus,

while in the present work very few predictions are presented, we expect that applying this method to more genomes will greatly enhance its power.

Discussion

We have shown that the expression levels of genes encoding interacting proteins tend to coevolve in yeast. This coevolution is of a fundamentally different nature than the only other type of coevolution that has thus far been studied in interacting proteins, namely the coevolution of amino acid sequence, and it may represent a widespread and important mode of evolutionary change. Both types of coevolution can be detected in scores of genes using a large set of protein interactions in yeast, though over three times more interacting pairs showed detectable coevolution of expression than of protein sequence in this study.

What is perhaps most surprising is the extent of coevolution we were able to detect using only four genome sequences. We did not use partial genome sequences that are available for many more yeast species (Cliften et al 2003; Souciet et al 2000), because including them dramatically reduced the number of genes for which alignments of orthologous genes in all species were available. However as many more yeast species will soon have complete genome sequences available, we expect that the power of the tests introduced here will increase greatly. Furthermore, our use of four genome sequences provides a reasonable benchmark for future studies in other eukaryotes such as *D. melanogaster*, *C. elegans*, and others, since close relatives of these species (*D. pseudoobscura* and *C. briggsae*) have already been fully sequenced and several other close relatives will soon have sequenced genomes. Our method may not be as easily

applicable in species with very little codon bias determined by gene expression levels, such as humans.

Aside from being useful for studying the evolution of gene regulation, our finding of expression coevolution has a practical application as well, in predicting pairs of interacting proteins. Since these predictions are more accurate when the expression coevolution metric is combined with another method of interaction prediction based on amino acid sequence coevolution, we suggest that future studies in which protein interactions are predicted from genome sequences will be more comprehensive if expression coevolution is included. Because even our combined metric cannot detect most protein interactions when only four genome sequences are used, we have not yet attempted to make large-scale predictions of interacting proteins in yeast.

In addition to the metric of expression coevolution that we introduce here, several other purely sequence-based methods for predicting protein interactions exist, such as phylogenetic profiling (Pellegrini et al 1999), conservation of gene neighborhood (Dandekar et al 1998), and gene fusions (Marcotte et al 1999; Enright et al 1999). Since these methods are mostly independent of one another, combining them might greatly increase our power to predict protein interactions based on genome sequences alone. The methods could be integrated in a Bayesian framework (as in Jansen et al 2003); for example, the extent of expression coevolution could serve as a prior probability of interaction, which can then be increased or decreased based on any other metric for interaction prediction. We note, however, that these other methods of protein interaction prediction would not have added any information in this study: phylogenetic profiling depends on the absence of some genes from some genomes, but all genes we used were

present in all four genomes; conservation of gene neighborhood requires shuffling of genes, but all genes we used had conserved synteny in the four genomes; and the method of gene fusions depends on relatively rare fusion events, which none of our genes have undergone in these four species.

Another unexplored application of both sequence and expression coevolution metrics is assessment of the quality of high-throughput protein interaction data sets (e.g. von Mering et al 2002). One could use the degree of expression and sequence coevolution in a set of putative protein interactions to determine how accurate the data are, using a set of well established interactions to determine a baseline of the maximum amount of coevolution expected to be seen if all interactions in a list were correct.

It is interesting to speculate about the possible future direction of work investigating expression coevolution. Current research into the cis-regulatory gene expression "code" of yeast, *Drosophila*, and other organisms may soon make it possible to predict the approximate expression patterns of genes in different conditions on a genome-wide scale (Beer and Tavazoie 2004). If this becomes possible, it will greatly increase the power to detect expression coevolution from sequence alone: instead of a single number (mean gene expression level, estimated by codon bias), one could calculate a vector representing the expression over many conditions for each gene in each organism. With this more detailed picture of gene expression regulation across different species, expression coevolution could be studied in far greater detail.

Finally, it is possible that coevolution of both protein sequences and expression levels may also be a property of pairs or groups of genes that do not necessarily interact physically. Larger groups, or modules, of genes that work together to produce some

output or trait (e.g., a single metabolic pathway) may show coordinated changes in expression levels and/or evolutionary rates due to increased or decreased utilization of those genes over evolutionary time. For example, if the genes specifically responsible for galactose transport and metabolism in yeast (the GAL genes) were used quite often in one species but seldom or never in another related yeast, we would expect to see an increase in the average expression (and thus codon bias) of those genes in the species that metabolized galactose more often. Changes in evolutionary rates might also be seen, since the species that seldom used galactose for energy would have little selective pressure to maintain the amino acid sequences of those genes; they would drift more than their orthologous counterparts in the other species, and this may be reflected as coevolution of amino acid sequences. Such coevolution at the levels of both expression and sequence evolution may allow inference of functional relationships between groups of genes that do not necessarily physically interact; this evolutionary approach to prediction of genetic relationships and functions may prove to be quite useful as the amount of genome sequence data continues to increase at an ever-faster rate.

Methods

Sequence data

For all analyses described in this work, we used the complete genome sequences of four closely related (<20 million years divergence, corresponding to an average of 2.2 synonymous substitutions/site after correcting for non-neutral synonymous sites) yeast species in the genus *Saccharomyces*: *S. cerevisiae* (Goffeau et al 1996), *S. paradoxus*, *S. mikatae*, and *S. bayanus* (Kellis et al 2003). Rigorous assignments of orthology were

made based on both high sequence identity and synteny between species (Kellis et al 2003), and alignments were done on protein sequences using ClustalW (Thompson et al 1994). Alignments were discarded if their maximum likelihood phylogeny (Yang 1997) was not consistent with the known phylogeny of the species, or if they contained either real or spurious (due to sequencing errors) frameshift mutations, since frameshifts result in unrealistic estimates of evolutionary rates. Frameshifts were detected by establishing a majority-rule consensus sequence from the four sequences; if any one sequence failed to match the consensus for at least five consecutive positions, it was counted as having a frameshift and discarded from the alignment.

Detection of sequence coevolution.

Our test for protein sequence coevolution of interacting proteins is similar to methods that search for strong correlations between pairwise sequence distances, or similarity of phylogenetic trees (Pazos et al 1997; Goh et al 2000; Ramani and Marcotte 2003; Pazos and Valencia 2001; ibid 2002; Goh and Cohen 2002). For each set of orthologous genes, we used PAML (Yang 1997) to estimate the evolutionary rate (dN/dS, or number of nonsynonymous substitutions per nonsynonymous site divided by synonymous substitutions per synonymous site) in each branch of the yeast phylogenetic tree. Five branch lengths were calculated for each set of orthologs (one for each of the four species plus one internal branch). These five lengths were normalized by dividing each by the average length of that branch over all the trees calculated, in order to control for the fact that some branches tended to be longer than others. The normalized lengths could then be plotted against each other for any pair of genes, and the Pearson correlation

coefficient (18) calculated as a measure of the degree of coevolution. To calculate the significance of the observed distribution of correlation coefficients among interacting pairs, we compared it to the distribution of all possible pairs except for those in the list of interactors. The nonparametric Kolmogorov-Smirnov test (KS test; Sokal and Rohlf 1995) was used to estimate the probability that both were sampled from the same underlying distribution.

Detection of expression coevolution

Our method for detecting expression coevolution was quite similar to our method for detecting sequence coevolution. Codon bias values, as represented by the codon adaptation index (CAI; Sharp and Li 1987), were calculated for each of the four orthologous sequences using the codon frequencies of the 20 most highly expressed genes in *S. cerevisiae*, as estimated by Arava et al. (2003). Results were not affected by using species-specific codon usage tables (not shown). The four values for each gene were then plotted against each other for each pair of genes, and the Pearson correlation coefficient was calculated for each pair. Details of significance testing by the KS test were as described above.

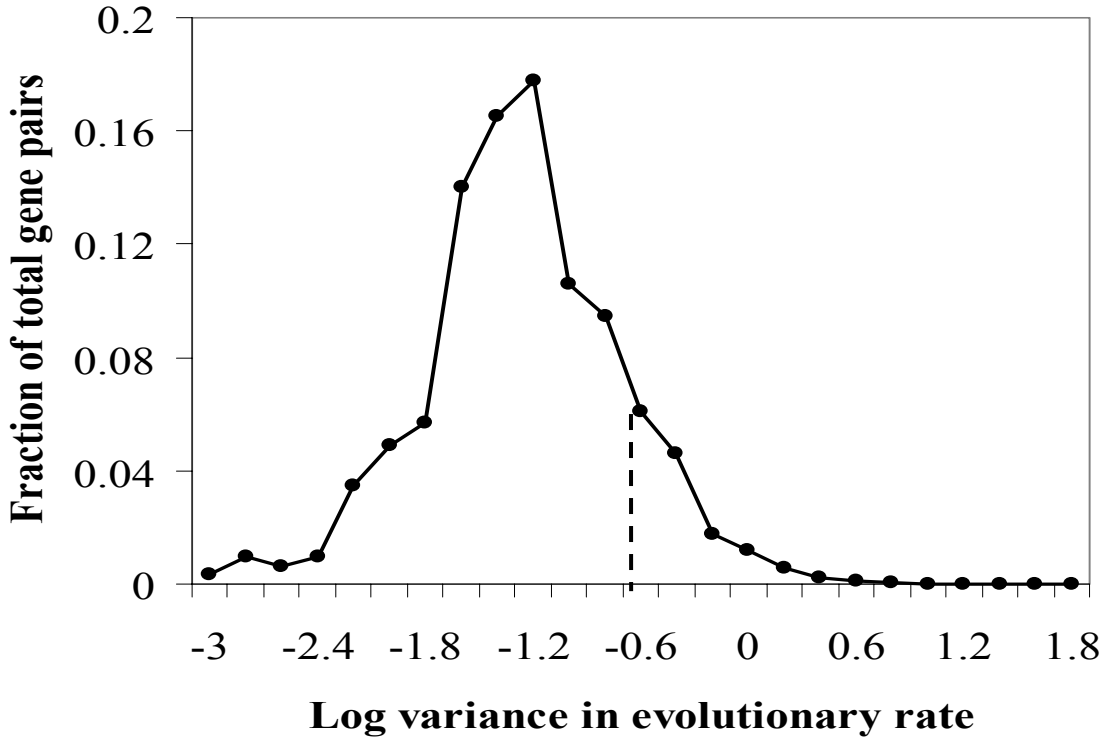
Protein-protein interaction data

A list of 4175 putative interactions involving 1360 *S. cerevisiae* proteins was taken from von Mering *et al*. (2002). Only those interactions listed with "high confidence" (interactions found by multiple independent methods) or listed as previously annotated (by non-high throughput methods) were used, in order to minimize the effects

of false positives. High-throughput methods used to identify interactions were yeast 2-hybrid, mass spectrometry, synthetic lethality, and synexpression; computational methods used were conserved gene neighborhood, gene fusion, and phylogenetic profiling (von Mering et al 2002). Exclusion of interactions whose membership in the "high confidence" category depended on synexpression (correlated expression levels in *S. cerevisiae* microarray experiments), because of a possible circularity when measuring CAI coevolution of these putatively interacting proteins, did not appreciably affect the results. Any interactions involving a protein with itself were discarded because these would show perfect coevolution for a trivial reason.

Figure 1.

Coevolution of sequence. A. A histogram of the base 10 logarithms of variance in evolutionary rates for all 192,510 possible pairs of proteins in this study. The variance for each protein in a pair was calculated, and the lower of the two was used to represent the pair. The dashed line indicates the variance cutoff described in the main text. Note that evolutionary rates were normalized by the mean rate for each branch of the phylogenetic tree (see Methods). B. A histogram of the correlation coefficients indicating the strength of amino acid sequence coevolution for 200 pairs of interacting proteins (solid line) and 26,596 pairs of non-interacting proteins (dashed line). The two distributions are significantly different from one another (KS test p=0.0069). Bin labels are the upper bound for each bin (e.g., the label 0.9 indicates $0.8<r\leq0.9$).


0.2
0.16
0.12
0.08
0.04
0
Fraction of total gene pairs
-3
-2.4
-1.8
-1.2
-0.6
0
0.6
1.2
1.8
Log variance in evolutionary rate

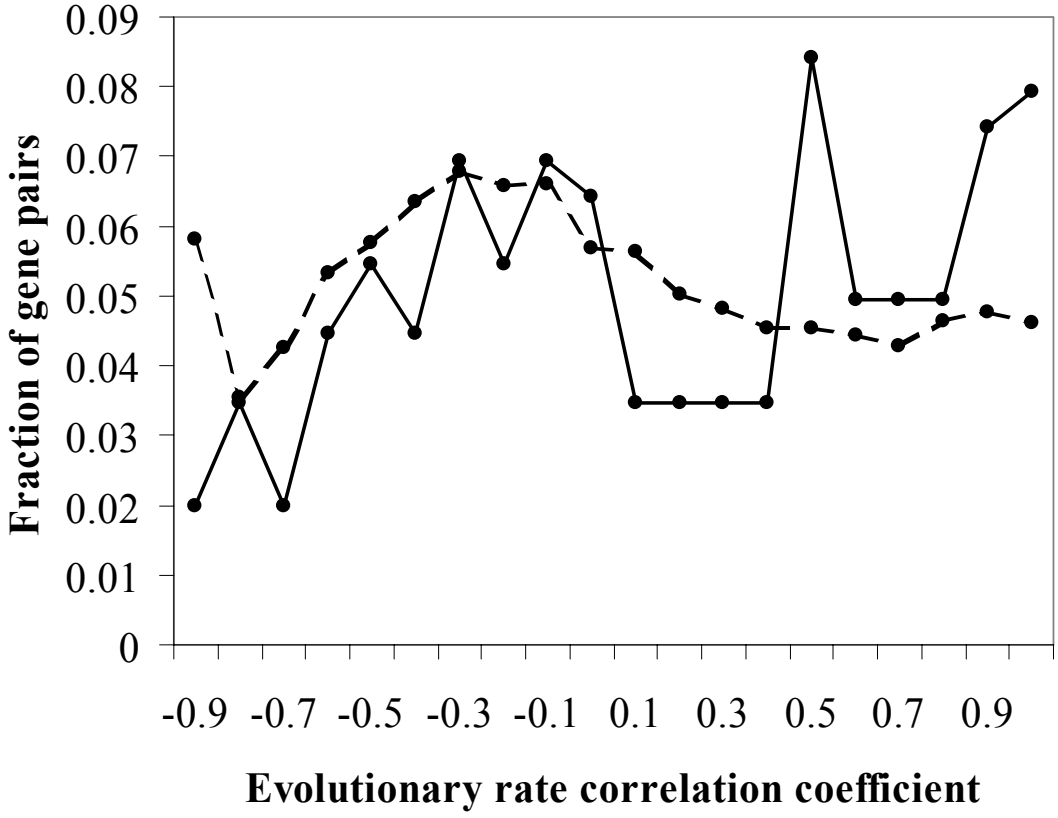
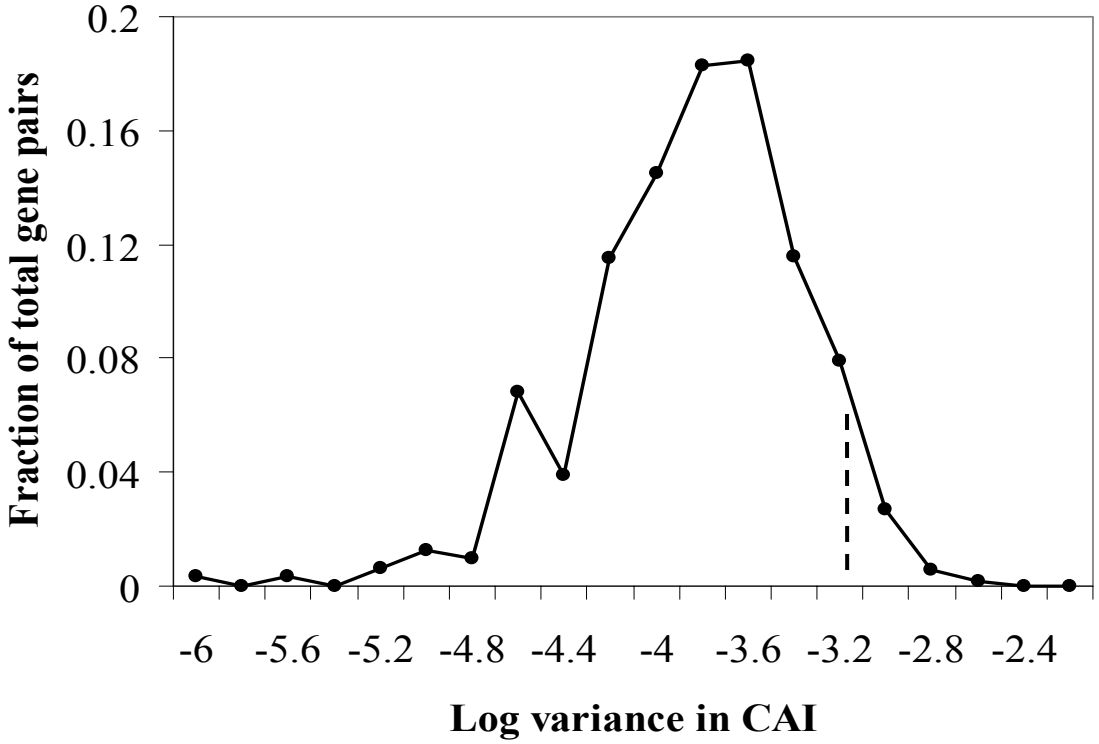

0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Fraction of gene pairs
-0.9
-0.7
-0.5
-0.3
-0.1
0.1
0.3
0.5
0.7
0.9
Evolutionary rate correlation coefficient

Figure 2.

Coevolution of expression. A. A histogram of the base 10 logarithms of variance in codon bias (CAI) for all 192,510 possible pairs of the 621 proteins in this study. The variance for each protein in a pair was calculated, and the lower of the two was used to represent the pair. The dashed line indicates the variance cutoff described in the main text. B. A histogram of the correlation coefficients indicating the strength of CAI coevolution for 200 pairs of interacting proteins (solid line) and 11,581 pairs of non-interacting proteins (dashed line). The two distributions are significantly different from one another (KS test $p<10^{-26}$). Bin labels are the upper bound for each bin (e.g., the label 0.9 indicates $0.8<r\leq0.9$).


0.2
0.16
0.12
0.08
0.04
0
Fraction of total gene pairs
-6
-5.6
-5.2
-4.8
-4.4
-4
-3.6
-3.2
-2.8
-2.4
Log variance in CAI

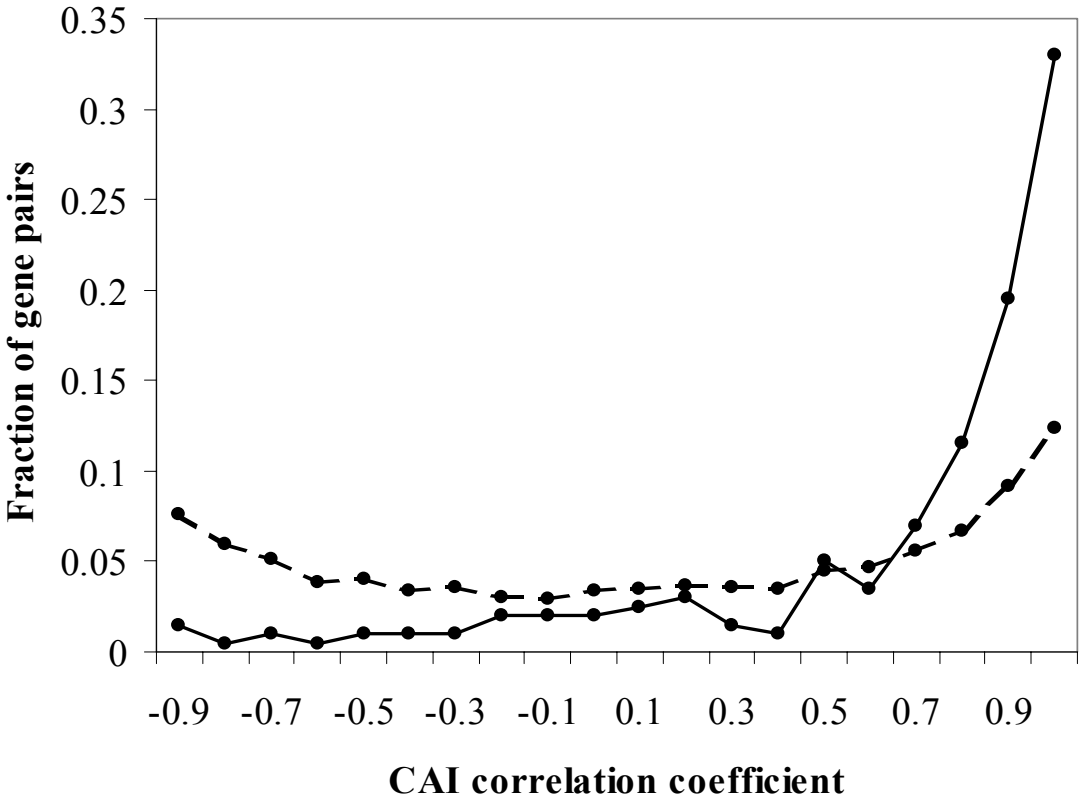

0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Fraction of gene pairs
-0.9
-0.7
-0.5
-0.3
-0.1
0.1
0.3
0.5
0.7
0.9
CAI correlation coefficient

Table 1.

Predictions of protein interactions. A list of 21 protein interaction predictions made by combining the sequence and expression coevolution metrics. The first ten pairs satisfy the stringent cutoffs of evolutionary rate $r>0.9$, CAI $r>0.95$; all 21 satisfy the cutoffs of evolutionary rate $r>0.75$, CAI $r>0.9$.

ORF 1	ORF 2	Known interaction?	CAI *r*	Evol rate *r*
FUR1	NOP7	NO	1.000	.999
RLP24	NOP7	NO	.962	.995
RLP24	FUR1	NO	.953	.991
MRPL33	RIB3	NO	.986	.944
PNO1	TIF35	NO	.995	.927
NOG1	RLP24	YES	.998	.920
SWP1	ATP3	NO	.980	.933
NOG1	FUR1	NO	.951	.948
NOG1	NOP7	YES	.960	.936
TIF2	YBR025C	YES	.967	.903
TAF17	QCR7	NO	.903	.969
VMA13	MLC1	NO	.904	.959
TIF2	RPL9A	NO	.988	.851
WBP1	YPT10	NO	.953	.844
RPL9A	YBR025C	NO	.945	.829
NOP7	UTP6	NO	.953	.813
FUR1	UTP6	NO	.962	.803
RPL5	YBR025C	NO	.901	.828
RPP0	TIF2	NO	.954	.761
RPL5	RPP0	YES	.936	.750
NOB1	APT1	NO	.912	.765

Chapter IV

Modularity and evolutionary constraints.

The majority of this chapter was previously published as Fraser HB, *Nature Genetics*, 37: 351 (2005)

On the role of modularity in evolution

Abstract

Modularity, which has been found in the functional and physical protein interaction networks of many organisms, has been postulated to affect both the mode and tempo of evolution. Here I show that in the yeast *Saccharomyces cerevisiae,* protein interaction hubs situated within single modules are highly constrained, while those connecting different modules are more plastic. This pattern of change could reflect a tendency for evolutionary innovations to occur by altering the proteins and interactions between rather than within modules, in a manner somewhat similar to the evolution of new proteins through the shuffling of conserved protein domains.

Introduction

Modularity in living systems has been a subject of inquiry for over 70 years (Needham 1933), though only recently has it become an object of intense study (Gerhart and Kirschner 1997; Hartwell et al 1999; Schlosser 2002; Schlosser and Wanger 2004). Similar to others, I define a module as a group of proteins that carries out a semi-autonomous function, which is often characterized by a higher density of functional linkages within modules than between them. Modularity is thought to increase evolvability (Gerhart and Kirschner 1997; Hartwell et al 1999; Schlosser 2002; Schlosser and Wanger 2004) both by providing a set of reusable "parts" that can be co-opted for new functions, and by reducing pleiotropy (i.e., the extent to which single genes affect

multiple traits) so that traits are able to be optimized by natural selection individually (Fiser 1930; Waxman and Peck 1998).

High-throughput protein-protein interaction screens have recently allowed the construction of networks consisting of thousands of interacting proteins (von Mering et al 2002); however higher-level attributes of these networks, such as their modular and temporal organization, are only beginning to be examined. A recent study found that proteins that physically interact with many other proteins ("hubs") can be classified into two largely distinct classes: those that appear to interact with most or all of their partners simultaneously ("party" hubs), and those that interact with different partners at different times ("date" hubs) (Han et al 2004). The study showed that the former type tends to connect proteins within functional modules, whereas the latter usually bridges different modules; for this reason, I will hereafter refer to party hubs as "intramodule" and date hubs as "intermodule". The two hub types also differ in their degree of pleiotropy: intermodule hubs are expected to be more pleiotropic, since they usually interact with multiple modules, in contrast to most intramodule hubs which only belong to a single module. This reasoning is supported by the finding that intermodule hubs have far more synthetic lethal interactions than intramodule hubs (Han et al 2004), reflecting their greater sensitivity to a wide range of genetic perturbations. This objective classification of proteins as intramodule or intermodule hubs is important, not least because it allows one to begin to address a fundamental question of evolution: how do genetic networks, and their constituent modules, evolve?

Results

One way to address this question is to compare the evolutionary constraints experienced by each hub type. At least three general scenarios are possible: first, it could be imagined that intermodule hubs evolve more slowly due to their higher pleiotropic constraints, consistent both with Fisher's idea that more pleiotropic mutations are less likely to be advantageous (Fisher 1930) and with more recent theoretical work (Waxman and Peck 1998). A second model would predict the opposite relationship: if modules often act as cohesive units that are not themselves very evolvable but instead must (once established) preserve their functions over vast stretches of evolutionary time, then species may be forced to evolve by altering the connections and interactions between rather than within modules (Hartwell et al 1999; Schlosser 2002), leading to greater constraint on intramodule hubs. A third model is that modularity has no relationship with evolutionary constraint, in which case (all else being equal) both types of hubs would be expected to experience similar levels of constraint.

To distinguish between these three models of network evolution, I used evolutionary rates (nonsynonymous to synonymous substitution ratio, or dN/dS) of yeast genes, calculated from four complete genomes of *Saccharomyces* sensu stricto species with a correction for non-neutral synonymous substitutions (Hirsh et al 2005). Strikingly, the mean evolutionary rate of the intramodule hubs was less than half that of the intermodule hubs (Fig 1a; fold difference=2.4), and the complete distributions of rates were significantly different (Kolmogorov-Smirnov [KS] test, $p<10^{-5}$). While both hub types evolve more slowly than proteins with no known interactions (as expected [Fraser et al 2002]), the difference was far more pronounced for intramodule hubs (Fig 1a; fold

difference=2.9, KS $p<10^{-18}$) than for intermodule hubs (Fig 1a; fold difference=1.2, KS p=0.002). Another way to demonstrate the difference between hub types is to assign each hub an arbitrary number based on its type (intramodule=1, intermodule=0), and then ask through correlation analysis, how well are evolutionary rates explained by hub type? The Spearman rank correlation coefficient for dN/dS *vs*. hub type was r=-0.44 ($p<10^{-7}$), confirming that hub type contributes significantly to evolutionary constraint. These results are consistent with the finding that human genes with many coexpression partners tend to be well conserved (Jordan et al 2004).

A related question is, are the phylogenetic distributions of the two hub types significantly different from one another? While this metric is not entirely independent of evolutionary rates, it nevertheless provides useful information about the approximate phylogenetic breadth of different proteins. Among seven diverse eukaryotic genomes (Krylov et al 2003), intramodule hubs are identifiable in significantly more species than are intermodule hubs (6.1/7 *vs*. 4.6/7; KS $p<10^{-4}$), consistent with the results found using dN/dS. Furthermore, 35.2% of intermodule hubs cannot be assigned to clusters of orthologous groups (Krylov et al 2003) (using the same seven genomes), compared to only 11.1% of intramodule hubs (Fisher's exact test $p<10^{-4}$).

To address the possibility that some information is being lost in assigning hubs to one of two types, I tested whether the primary metric used to distinguish between hub types (the average expression correlation between the hub and its interactors across many microarray experiments, or strength of coexpression; see Supplementary Notes later in this chapter) contains more information than the discrete classification of hub types that is largely based upon it. The strength of coexpression is even more strongly correlated

with dN/dS than is the hub type classification (Fig. 1b; Spearman r=-0.57, $p<10^{-14}$), explaining 32.5% of the variance in evolutionary rates. The eukaryotic phylogenetic breadth (Krylov et al 2003) of hubs is also correlated with their strength of coexpression (Spearman r=0.43, $p<10^{-11}$). These results further suggest that intramodule hubs evolve more slowly, and are present in more eukaryotic species, than are intermodule hubs.

Because many factors are known to influence evolutionary rates of proteins, and these could possibly lead to indirect correlations, I tested the association of hub types with several possible confounding factors. These include fitness effect (or "gene dispensability"), number of protein-protein interactions, mRNA expression level, type of interaction (stable or transient), and functional class. Of all these factors, only expression level plays more than a trivial role, and even this variable does not explain as much of the variance in evolutionary rates as does coexpression strength (see Supplementary Notes later in this chapter).

It can be concluded from this analysis that modularity affects the evolutionary constraints on proteins, in support of the second model of genetic network evolution presented above (Hartwell et al 1999; Schlosser 2002): central constituents of modules are highly constrained in their evolution, whereas the proteins connecting different modules are more labile. This modular organization explains about one third of the variance in evolutionary rates of these proteins, more than any other factor reported to date, emphasizing the key role modularity appears to play in evolution (for a more detailed discussion, see Supplementary Notes later in this chapter).

In an abstract sense, this result parallels the idea of evolutionary innovation occurring through exon shuffling (Doolittle 1995), a process in which the swapping

between different genes of DNA segments encoding conserved protein domains allows new proteins to evolve, by creating novel domain combinations (the fact that domains can be found in a multitude of different combinations [Doolittle 1995] is strong evidence that this process occurs). The analogy here is that functional modules may be similar to protein domains, which once established are usually best left intact; new combinations of and connections between modules (or domains) may more often lead to advantageous changes than would changes that disrupt modules (or domains). This model is, in some sense, similar to the idea of evolution occurring through the co-option of existing modules for new purposes (Gerhart and Kirschner 1997; Hartwell et al 1999; Schlosser 2002; Schlosser and Wanger 2004). One prediction of this model is that intermodule hubs should gain and/or lose interactions faster than intramodule hubs; this should be testable once reliable large-scale protein interaction data are available for another yeast, such as *Candida albicans*.

It is quite likely that modularity affects many more aspects of evolution than have been discussed in this work. For example, protein sequences and expression patterns may coevolve within modules, since alterations in the sequence or expression of one protein can result in a selection pressure for reciprocal changes in other members of the same module (Schlosser 2002; Fraser et al 2004). As the number of fully sequenced genomes continues to increase at an ever-faster rate, such coevolution should become trivial to identify, and it will be interesting to see if this and other predictions of the effects of modularity on evolution are borne out.

Supplementary notes on modularity and evolution

Controlling for possible confounding variables

As described in the main text, I controlled for a number of possible confounding variables that could lead to an indirect correlation between coexpression strength (or hub type) and evolutionary rates. First I tested three variables known to correlate with selective constraint: growth rate of yeast when a gene is deleted (or fitness effect [Hirsh and Fraser 2001]), number of protein-protein interactions (Fraser et al 2002), and mRNA expression level (Pal et al 2001). Following Han et al. (2004), I excluded ribosomal subunits from the following analyses, because they represent outliers with very high coexpression strength and very low evolutionary rates (as expected, excluding ribosomal subunits weakened the correlation between coexpression strength and evolutionary rate; Spearman r=-0.50). Consistent with Han et al.'s (2004) finding that gene essentiality was not different between the hub types, neither were fitness effects (Giaever et al 2002; Steinmetx et al 2002) (data not shown); thus controlling for fitness effects or gene essentiality has no effect on the correlations. The number of protein-protein interactions (as reported by Han *et al*. [2004]) was different between the two hub types, being significantly (Kolmogorov-Smirnov [KS] test [Sokal and Rohlf 1995] $p<10^{-3}$) more numerous for intermodule hubs (mean interactions = 10.2) than for the intramodule hubs (mean interactions = 8.4). However because more protein interactions are known to correlate with slower evolutionary rate (Fraser et al 2002), this actually weakens the apparent effect of hub type on evolutionary rate. Accordingly, statistically controlling for number of interactions using the method of partial correlation (which allows one to examine the correlation between two variables while controlling for a third, potentially

related variable [Sokal and Rohlf 1995]) increases the correlation between coexpression strength and evolutionary rate, though only very slightly. Lastly, gene expression levels (Wang et al 2002) were significantly correlated with hub type (Spearman r=0.37, $p<10^{-7}$), indicating that intramodule hub mRNAs are more highly expressed in log-phase rich glucose medium growth. In order to separate the effects of gene expression and hub type on dN/dS, the partial correlation method was again employed. The correlation between coexpression strength and dN/dS was still quite significant after controlling for expression level (Spearman r=-0.32, $p<10^{-5}$), and was slightly stronger than that between expression level and dN/dS controlling for coexpression strength (Spearman r=0.28, $p<10^{-4}$), indicating that coexpression strength explains more of the variance in dN/dS than does expression level for these genes. (It is worth noting that while controlling for coexpression strength weakens the correlations of these other variables with dN/dS among the protein interaction hubs, it is not possible to tell what effect controlling for coexpression strength would have on any genome-wide correlations, since only protein interaction hubs are being studied here.)

Another possible confounding factor is the kinds of protein interactions in which the two hub types participate. As reported by Han et al. (2004), according to the YPD database (Costanzo et al 2001) 81% of the intramodule hubs participate in stable protein complexes, compared to only 19% of the intermodule hubs. Because it is possible that protein complex subunits experience different evolutionary pressures than other proteins (Teichmann 2002), I controlled for this by comparing the mean dN/dS of the two hub types, using protein complex subunit hubs only (examination of high-throughput mass spectrometry protein complex data (Gavin et al 2002; Ho et al 2002) revealed that many

of the hubs classified as non-complex subunits may in fact participate in stable complexes, but for this analysis the more conservative list of complex subunits from the YPD database was used). The two hub types actually showed an even greater difference in evolutionary rates among complex subunits only (ratio of dN/dS means = 3.1, KS $p<0.002$; the weaker p-value is due to the smaller sample size), demonstrating that the difference in dN/dS is not due to disparate numbers of protein complex subunits in the two hub types. Also, there is still a highly significant correlation between coexpression strength and dN/dS when examining only protein complex subunits (Spearman $r=-0.40$) that is only slightly weaker than the overall correlation. Comparing the mean dN/dS values for only non-complex subunit hubs did not yield a significant result, but this is most likely due to having dN/dS values for only nine intramodule hubs not listed as complex subunits by Han et al. (2004) (and as stated above, even among these intramodule hubs classified as non-complex subunits are some likely complex subunits, as indicated by high-throughput mass spectrometry studies [Gavin et al 2002; Ho et al 2002]).

Yet another difference that could influence evolutionary rates in the two hub types is their enrichment for proteins of different functional categories. For example, the intramodule hubs are enriched for proteasome subunits (16 genes), transport ATPases (9 genes), and RNA polymerase III subunits (9 genes) (as mentioned above, ribosomal subunits have already been excluded from the analysis). It could be imagined that some functional class, by its enrichment in one hub type and its slow evolution, could significantly influence the mean evolutionary rate of that hub type. If this were the case, the results presented above could simply demonstrate that intramodule hubs are enriched

for some functional class and that members of this class evolve slowly; it would not be evidence for a causal relationship between hub type and evolutionary rate (though such a causal relationship would still be quite possible; as an extreme example, if all intramodule hubs were of a single slowly evolving functional class, then it could still be that the reason this class evolves slowly is because of its enrichment for intramodule hubs, instead of the intramodule hubs evolving slowly because of their enrichment for this functional class). To address this possibility, I repeated the comparison of dN/dS for the two hub types, excluding different functional classes enriched in one of the two types. No single class had a large effect on the results; the strongest effect was found by exclusion of the proteasome subunits from the intramodule hubs, which only reduced the fold difference in mean dN/dS between the two classes from 2.4 to 2.2, still a highly significant difference (KS $p<10^{-4}$). Similarly, elimination of the most over-represented class in the intermodule hubs (13 SAGA/ADA complex members) did not weaken the difference in evolutionary rates; in fact, it increased the disparity (mean dN/dS ratio = 2.6; KS $p<10^{-5}$), and no other functional class over-represented in the intermodule hubs significantly contributed to the difference in evolutionary rates. These results, as well as the finding that fitness effects of the two hub types do not differ from one another (and thus neither type is differentially enriched for functional classes that are more or less important, for general growth and fitness of yeast, than classes in the other hub type), indicate that while each hub type is enriched for different functional classes, no such class contributes disproportionately to the observed difference in evolutionary constraint.

Discussion of the relationship between protein connectivity and essentiality

Another implication of these results concerns the relationship between network disruption and protein dispensability. Proteins with many physical interactions have been found to have both greater fitness effects (Fraser et al 2002) (among non-essential genes) and a greater chance of being essential (Jeong et al 2001) than those with few interactions; the explanation for this relationship proposed by Jeong *et al*. (2001) and Barabasi and Oltvai (2004) was that highly connected proteins have a greater negative effect on the network diameter (defined as the average shortest path length between two proteins) when removed than do less connected proteins. The set of interaction hubs studied here provide an opportunity to test this hypothesis, since the removal of intermodule hubs has a far greater effect on network diameter than does the removal of intramodule hubs (Han et al 2004). The fact that there is no significant difference in the essentiality or fitness effects of these two groups indicates that a protein's contribution to network diameter is most likely not the explanation for the increased essentiality of highly interactive proteins (Jeong et al 2001), since if it were then we would expect many more intermodule hubs to be essential for viability. Indeed, the finding that intramodule hubs are the more evolutionarily constrained class indicates that a protein's contribution to network diameter is also most likely not relevant to evolutionary constraint in any straightforward manner, since if it were then one would expect to see the opposite relationship with evolutionary rates. It should be noted that the hubs under consideration here are not a representative subset of all genes, so this analysis does not exclude the possibility that network diameter contribution is relevant to the essentiality of some non-

hub proteins; however it does indicate that this factor is unlikely to be able to account for the general relationship between number of interactions and essentiality.

Caveats concerning the interpretation of these results

It is worth noting that the general model I have proposed to explain the difference in evolutionary constraint between different hub classes does not require modularity to have evolved in order to promote evolvability. The difference in evolutionary rates would simply reflect the history of changes in the yeast lineage, and the tendency for changes in some proteins to be more tolerable (or advantageous) than those in others. However, while evolvability is not a necessary component of this mechanism, at the same time its role cannot be excluded; modules may often act as functional cassettes that can promote evolvability by alteration of their intermodule connections leading to co-option for a new purpose. Such co-option is well documented for some modules, such as the *hedgehog*/*patched* developmental module, which is used in a variety of contexts in metazoans (True and Carroll 2002; von Dassow and Munro 1999). Even if modularity can be shown to increase evolvability, though, this does not mean it was selected to do so.

While the results presented here do not provide support for one prediction that can be made from the theory that pleiotropy decreases evolvability, three caveats are in order. First, these results do not exclude the possibility that more pleiotropic genes do evolve more slowly than less pleiotropic genes in general, but that in the set of genes examined here this effect is overpowered by the effects of each protein's position within or between modules. A rigorous test of this would require a reliable method for quantitatively

measuring the degree of pleiotropy for many genes, which unfortunately is not presently possible. Indeed, we do not even possess a sufficient understanding of pleiotropy to know how to capture a concept as complex (Hodgkin 1998) as pleiotropy in a single number for each protein (For example, is a protein with several catalytic functions more or less pleiotropic than a protein with one catalytic function that is used in several different pathways?). For this reason, my use of each protein's number of synthetic lethal interactions as reflective of degree of pleiotropy (to lend support to the theoretical expectation for intermodule hubs to be more pleiotropic) should be seen only as a rough approximation.

Second, these findings by no means speak against the general idea that pleiotropy reduces organismal (or species) evolvability, since the effects of pleiotropy on the evolutionary rates of individual proteins most likely do not have any simple relationship with the effects of pleiotropy on organismal evolvability. The results presented here begin to address several questions regarding modularity and constraint on proteins, but not modularity and evolvability of entire organisms; to test the latter would require (for example) comparing the evolvability of two lineages identical in every respect except for the average degree of pleiotropy of their genes. While this is clearly not possible, information might still be gleaned from lineages that approximate this ideal (Yang 2001), though caution must be taken in the interpretation of such studies.

Third, while the different locations of the two hub types with respect to functional modules in the yeast protein interaction network provide a plausible explanation for the observed difference in evolutionary rates, other possibilities should be considered as well. Even though I was able to control for other factors known to influence evolutionary rates

such as fitness effect, number of protein-protein interactions, expression level, protein complex membership, and functional class, it remains possible that a previously unrecognized factor could be playing a role. For example, it could be that a protein with many simultaneous interactions experiences greater constraint on its structure than a protein with an equal number of sequential interactions, aside from considerations of modularity. However without a compelling mechanistic explanation for why this (or any other factor aside from modularity) might explain the differences in evolutionary rates of the two hub types, the most parsimonious explanation is that modularity plays an important role in determining evolutionary constraints on proteins.

Methods

Sources of data

Evolutionary rates of *S. cerevisiae* genes were calculated as previously described (Fraser et al 2004), with a correction for the effect of codon bias on dS (Hirsh et al 2005). The codeml program was used to calculate the maximum likelihood rates (Yang 1997), and the rate was allowed to vary across branches of the phylogenetic tree (model 1); only the *S. cerevisiae* branch rate was used in this work. dN/dS values were able to be calculated for 147 of the 199 non-ribosomal hubs (ribosomal proteins were excluded from the hub type classification by Han *et al*. (2004), so they are likewise not included in any analyses comparing intermodule and intramodule hubs, though they are included in Fig 1b because no hub type classification is used in this figure), after discarding alignments with any ambiguous orthology assignments, putative frameshifts, poor alignments, or missing sequences (Fraser et al 2004). Functional genomic data were taken

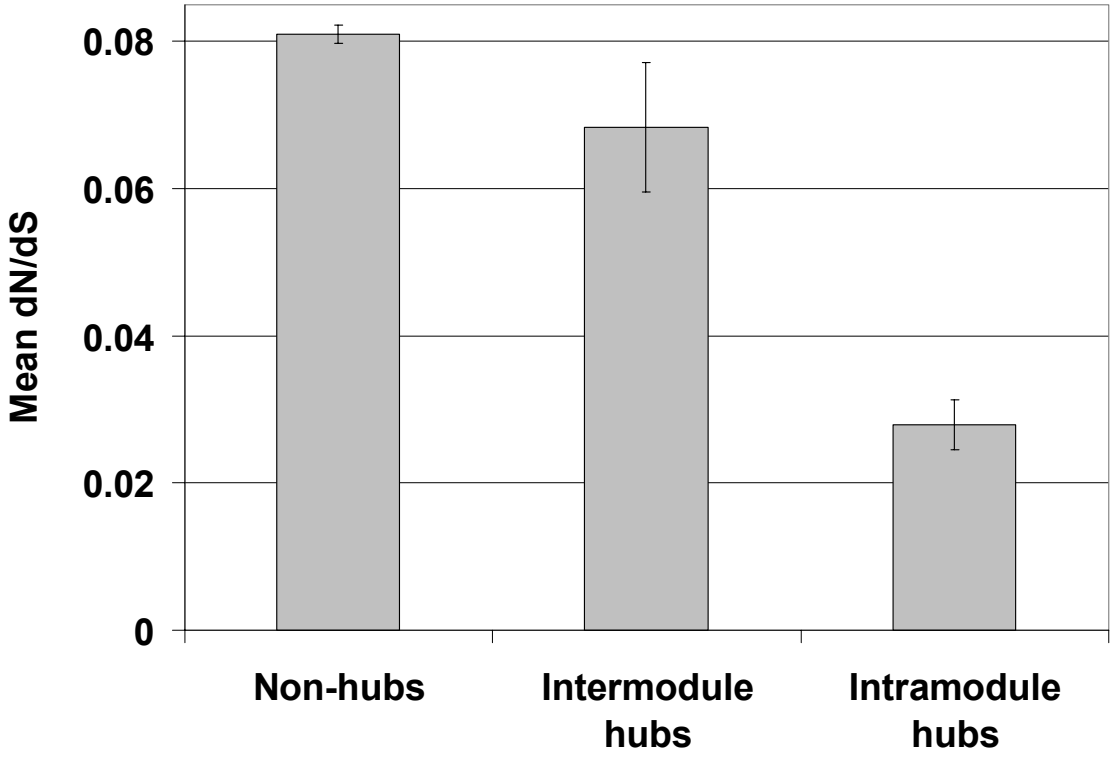
from references as cited in the section entitled "Controlling for possible confounding variables". For the phylogenetic distribution analyses, phylogenetic breadth was defined as the number of species (out of seven) in each cluster of orthologous groups (Krylov et al 2003). Genes from *S. cerevisiae* not present in any such clusters were classified as *S. cerevisiae*-specific (that is, not present in any of the other six species examined [Krylov et al 2003]); while these genes will not always be truly absent in all of the other six genomes, this classification should provide a roughly accurate picture of the phylogenetic distributions for all yeast genes.

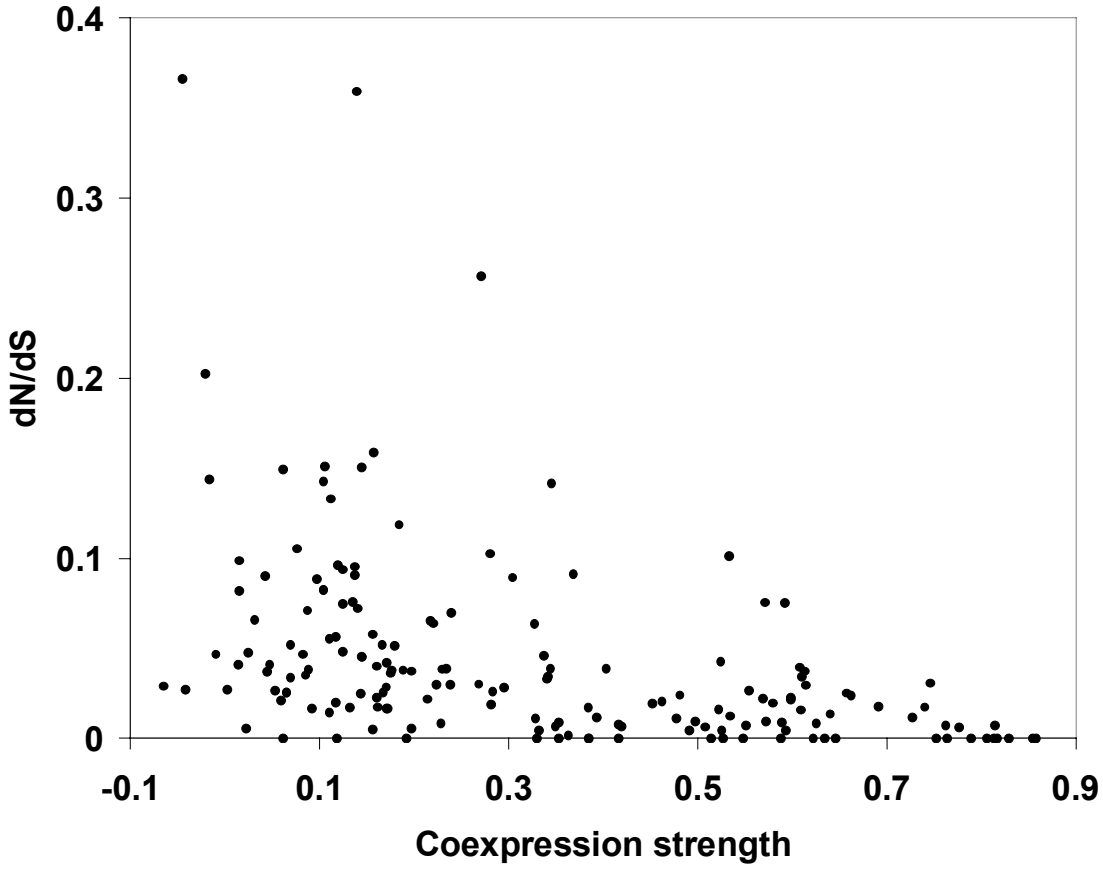
Coexpression strength and hub type

Han et al. (2004) classified a protein as a party hub if its average Pearson correlation coefficient with at least six interactors was above a threshold in either all microarray data sets together, or any one of five subsets of the data (stress response, cell cycle, pheromone treatment, unfolded protein response, and sporulation). Because these criteria cannot be captured in a single number (there are six different average correlation coefficients being examined), I used only the average correlation over the entire set of experiments as a metric of coexpression strength. Testing of other possible metrics, such as the maximum of the six average correlation coefficients, did not improve results.

Figure 1.

Intramodule hubs are more evolutionarily constrained than intermodule hubs. (a) Mean dN/dS values for three protein classes (± s.e.); intramodule hubs are more constrained than intermodule hubs, which are in turn slightly more constrained than proteins lacking any known physical interactions. (b) dN/dS values plotted against coexpression strength for all hubs. There is a negative correlation (Spearman r=-0.57, $p<10^{-14}$), indicating that hubs with more interacting protein coexpression (intramodule hubs) are more constrained than those with little coexpression (intermodule hubs).


0.08
0.06
0.04
0.02
0
Mean dN/dS
Non-hubs
Intermodule hubs
Intramodule hubs


0.4
0.3
0.2
0.1
0
dN/dS
-0.1
0.1
0.3
0.5
0.7
0.9
Coexpression strength

Chapter V

Evolution of gene expression noise minimization.

The majority of this chapter was published by Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. *Public Library of Science Biology* 2: e137 (2004).

Abstract

All organisms have elaborate mechanisms to control rates of protein production. However, protein production is also subject to stochastic fluctuations, or "noise". Several recent studies in *S. cerevisiae* and *E. coli* have investigated the relationship between transcription and translation rates and stochastic fluctuations in protein levels, or more generally how such randomness is a function of intrinsic and extrinsic factors. However the fundamental question of whether stochasticity in protein expression is generally biologically relevant has not been addressed, and it remains unknown whether random noise in the protein production rate of most genes significantly affects the fitness of any organism. We propose that organisms should be particularly sensitive to variation in the protein levels of two classes of genes: genes whose deletion is lethal to the organism and genes encoding subunits of multiprotein complexes. Using an experimentally verified model of stochastic gene expression in *S. cerevisiae*, we estimate the noise in protein production for nearly every yeast gene, and confirm our prediction that the production of essential and complex-forming proteins involves lower levels of noise than does the production of most other genes. Our results support the hypothesis that noise in gene expression is a biologically important variable, is generally detrimental to organismal fitness, and is subject to natural selection.

Introduction

Stochasticity is a ubiquitous characteristic of life. Such apparent randomness, or "noise", can be observed in a wide range of organisms, resulting in phenomena ranging from progressive loss of cell cycle synchronization in an initially synchronized

population of microbes to the pattern of hair coloration in female calico cats. An important source of stochasticity in biological systems is the random noise of transcription and translation, which can result in very different rates of synthesis of a specific protein in genetically identical cells in identical environments (Ozbudak et al. 2002, Elowitz et al. 2002, Blake et al. 2003).

Understanding how stochasticity contributes to cellular phenotypes is important to developing a more complete picture of how cells work. Accordingly, noise in gene expression and other cellular processes has been a major focus of research for more than a decade. While several cases where stochasticity is advantageous have been described (e.g., phase variation in bacteria [Hallet 2001] or the lysis/lysogeny decision in phage lambda [Arkin et al. 1998]), it is expected that noise is not advantageous in most cellular processes, as precisely controlled levels of gene expression are presumably optimal (e.g., Barkai and Leibler 2000). However, whether noise in expression is of consequence to organismal fitness has not previously been investigated, despite the centrality of this question to our understanding of the role of noise in biological systems.

In this study, we investigate whether the differences in noise levels among genes are consistent with the hypothesis that noise in gene expression has been subject to natural selection to reduce its deleterious effects. We propose that random fluctuations in the expression levels of two groups of genes in yeast, essential genes and protein complex subunits, should be particularly consequential for organismal fitness. If noise in gene expression is not an important factor to yeast—that is, if the level of stochasticity encountered by yeast in the expression of its genes is below that which would have negative consequences—then we would expect to see no difference in the randomness of

expression in genes for which noisy expression is predicted to be relatively more or less deleterious. However, if stochasticity is an important variable on which natural selection has acted, we would expect to see the strongest signature of such selection in the expression of genes for which yeast are the most sensitive to randomness.

Results

If deletion of a gene has only a small deleterious effect on the fitness of yeast, then random fluctuations in the amount of protein produced from that gene are likely to have a similarly small, or smaller, impact. In contrast, the same fluctuations in the level of a protein essential for viability may have a profound effect on fitness; in the extreme, fluctuation to levels below that required for normal cellular function could compromise viability. Considering this predicted difference in the sensitivity of yeast to randomness in expression of essential versus relatively dispensable genes, we reasoned that if noise in gene expression is a biologically important variable, selection for reduction of stochasticity in expression levels would likely be stronger for essential genes than for nonessential ones.

A recent study linking noise in protein levels to transcription and translation rates in yeast (Blake et al. 2003) allows us to test this prediction. In the study, noise in the expression of a GFP reporter gene was measured by flow cytometry; stochasticity was measured as the amount of variation in GFP levels per cell in a population. Thus if all cells in a population had very similar levels of GFP, there was little noise in the production of the GFP. The effect of transcription and translation on noise levels were studied by independently varying these two parameters, and measuring the resulting noise

levels for a population of cells. This experimental approach, as well as a mathematical model of protein production (Blake et al. 2003), indicates that noise in protein production is maximized at intermediate levels of transcription (at ~1/3 maximal transcription rate of a gene, regardless of what the maximal transcription rate is [Footnote 1]), as well as at maximal levels of translation per mRNA molecule.

To produce a given number of any particular protein, yeast could adopt one of three qualitatively different strategies (Fig. 1): 1. maximize transcription and minimize translation per mRNA; 2. maximize translation per mRNA and minimize transcription; or 3. employ intermediate levels of both transcription and translation per mRNA. Importantly, strategy 1 should result in less stochasticity than strategy 2 or 3. Strategy 2 is noisy due to the high translation, and strategy 3 is noisy due to both intermediate transcription and translation. In contrast, noise is minimized at both transcription and translation steps for genes that exhibit strategy 1. Thus we predicted that if noise in protein production is an important factor to yeast, then genes that are essential for viability would be biased towards having high transcription rates and a low number of translations per mRNA. The data currently available do not allow us to predict whether expression strategy 2 is more or less noisy than strategy 3.

The overall correlation between a gene's dispensability (defined as the growth defect of a yeast strain missing that gene in rich glucose medium; i.e., an essential gene is indispensable) and its rate of protein production (Supplementary Figure 1) presents a potentially confounding relationship in our analysis. Dispensability is correlated with the rate of protein synthesis for reasons that may have nothing to do with stochasticity, simply because most essential proteins are needed in somewhat greater quantity than

most nonessential proteins, so their genes must be more highly transcribed and/or translated. This correlation might lead to an association between gene importance and the likelihood of adopting expression strategy 1. In order to control for the overall correlation between dispensability and the rate of protein production, we employed two statistical methods.

In the first of these two methods, we binned yeast genes by their protein production rate (proteins/sec), so that all genes in each of 15 bins had approximately equal levels of protein production (see Supplementary Table 1 for details). These genes could have achieved this level by any of the three possibilities listed above; our prediction was that if noise in gene expression is relevant to yeast, then essential genes would be biased towards having the highest transcription and lowest translation per mRNA (strategy 1) in each bin. Indeed, this was confirmed by the data: when the genes within each bin were separated into thirds by their number of translations per mRNA, a larger number of essential genes were in the third with the lowest number of translations (low noise) than in the third with the highest number of translations (high noise) for all but one of 15 bins (Fig. 2a). A Fisher's exact test (Sokal and Rohlf 1995) demonstrated that all 14 bins with more essential genes in the low noise third than the high noise third had significantly more essential genes in the low-noise third ($P \leq 0.02$). Similar results were found when using different numbers of bins, using halves or quartiles instead of thirds, or when separating bins into by transcription rate instead of number of translations per mRNA (not shown). This result cannot be explained by the overall positive correlation between dispensability and rate of protein synthesis (Footnote 2).

Because binning genes still allows for a small amount of variability in protein production within each bin (Supplementary Table 1), we sought to control for protein production rate in another fashion as well. We employed partial correlation, a method that allows one to examine the relationship between two variables when other, possibly confounding, variables are statistically held constant (see Methods). The stochastic model of gene expression (Blake et al. 2003) led us to the prediction that, when protein production rate is controlled for, fitness effect (f, where f = 0 indicates no effect on growth when a gene is deleted, f = 1 indicates that a gene is essential, and $0 < f < 1$ indicates a quantitative growth defect [Hirsh and Fraser 2001]) would correlate positively with transcription rate and negatively with translation rate per mRNA. Indeed, this is what we observed (f vs txn rate | protein prod rate, Spearman partial r=0.282, n=4746, P=10^{-87}; f vs tlns per mRNA | protein prod rate, Spearman partial r=-0.258, n=4746, P=10^{-75}). We would also expect that the relationship between gene importance and implementation of the expression strategy that minimizes noise could additionally be seen by considering transcription rate and translation rate per mRNA together, as a ratio; a large txn rate/tlns per mRNA indicates transcripts are produced quickly but are translated slowly, corresponding to our expression strategy 1. Confirming this, the correlation between fitness effect and the ratio of transcription rate/translations per mRNA, while controlling for proteins production rate, is highly significant (f vs txn rate/tlns per mRNA | protein prod rate, Spearman partial r=0.275, n=4746, P=10^{-86}). Partial correlation analysis is thus in accordance with the trend illustrated in Fig. 2a: essential genes preferentially use expression strategy 1, which minimizes stochasticity.

In addition to essential genes, genes whose protein products participate in stable protein complexes ("complex subunits") would also be expected to exhibit sensitivity to randomness in expression: producing too little or too much of a single protein complex subunit can compromise the proper assembly of the entire complex, and waste the energy invested in production of the other complex subunits. In support of this, it has been found that both under- and over-expression of complex subunits is more likely to result in reduced growth rate or inviability of yeast than is misexpression of other genes, and also that complex subunits tend to be more precisely coexpressed with other genes than non-complex subunits (Papp et al. 2003). Using data from two high-throughput studies that identified proteins involved in stable complexes (Gavin et al. 2002, Ho et al. 2002), we assigned genes to two groups: those whose protein products are members of a stable complex found in either study, and those whose products are not (since the protein complex data do not include all protein complexes, we expect that many protein complex subunits will not be classified as such in our list; this, as well as any false positives in the data, results in weakening of our results and an underestimate the true strength of the effect). We then performed the same binning analysis as described above, substituting our list of complex subunits for our list of essential genes. Again the prediction was confirmed: in all 15 bins, the third of the bin with the least translation per mRNA (and thus the lowest noise level) contained more complex subunits than the third with the most translation per mRNA (Fig. 2b). The association between low translation per mRNA and protein complex membership was significant (Fisher's exact test $P \leq 0.02$) for all but one bin. As in Fig. 2a, this result is robust with respect to the number of bins and size of divisions within bins used (not shown). Also as in Fig. 2a, the bias is the opposite of that

expected from the positive correlation between fitness effect and protein production rate; it is also the opposite of what would be the result of the fact that highly transcribed genes are more likely to appear in the list of protein complex subunits than are poorly transcribed genes (Footnote 2).

When we repeated the partial correlation analysis for complex subunits (genes were assigned a value of 1 if they were a complex subunit, 0 if not), we found similar results. When total protein synthesis was controlled for with the partial correlation, complex subunits were more likely to have a high transcription rate (complex subunit vs. txn rate | protein prod rate, Spearman partial r=0.203, n=4900, P=10^{-46}) and a low number of translations per mRNA (complex subunit vs tlns per mRNA | protein prod rate, Spearman partial r=-0.200, n=4900, P=10^{-46}). Using the ratio of transcription rate to translations per mRNA also yielded similar results (complex subunit vs txn rate/tlns per mRNA | protein prod rate, Spearman partial r=0.220, n=4900, P=10^{-56}). Thus, partial correlations confirm the finding illustrated in Fig. 2b.

Since proteins that participate in many protein-protein interactions are more likely to be essential (Jeong et al. 2001, Fraser et al. 2002), it was not immediately clear if protein fitness effect and membership in a multiprotein complex are independently associated with the expression strategy that minimizes stochastic fluctuations. To address this question, we calculated the partial correlation between fitness effect and the ratio of transcription rate/translations per mRNA, while controlling for both protein production rate and protein complex membership. Likewise, we calculated the correlation between protein complex membership and transcription rate/translation rate per mRNA while controlling for both protein production rate and fitness effect. The two partial

correlations were both quite significant (f vs txn rate/tlns per mRNA | protein prod rate, complex membership: Spearman partial r=0.227, n=4746, P=10^{-57}; complex membership vs txn rate/tlns per mRNA | protein prod rate, f: Spearman partial r=0.147, n=4746, P=10^{-24}), suggesting that fitness effect and protein complex membership are independently associated with the expression strategy that minimizes stochastic fluctuation (however the relative strengths of the partial correlations cannot be interpreted as their true relative contributions, due to the differing quality of fitness effect and protein complex membership data). Repeating the partial correlations above with either transcription rate or translations per mRNA in place of their ratio gave significant partial correlations with both fitness effect and protein complex membership as well (not shown).

The hypothesis that genes of large fitness effect are under stronger selection to reduce stochastic fluctuation in expression provides an explanation for a previously observed, but as yet unexplained, correlate of protein evolutionary rate. Pal et al. (2001) noted a weak but significant negative correlation (r = -0.11, P = 10^{-9}) between an mRNA's rate of decay and the evolutionary rate of the protein it encodes. This correlation was surprising, as it is precisely the opposite of what one would expect, were the relationship between the rates of mRNA decay and protein evolution mediated by the level of expression: slow decay would result in increased expression, which is known to be associated with slow evolution (Pal et al. 2001). Thus we would expect a positive correlation between rates of mRNA decay and protein evolution; not the negative one that is observed. However, under the present hypothesis that relatively important genes are under stronger selection to reduce noise, the relationship between mRNA decay and protein evolutionary rate is interpretable. Both genes of large fitness effect and protein

complex subunits are known to evolve slowly (Hirsh and Fraser 2001, Jordan et al. 2002, Fraser et al. 2002; while the reason why genes of large fitness effect evolve slowly has been debated [Pal et al. 2002, Hirsh and Fraser 2002], the presence of the correlation has not been disputed, and it can be seen to be much stronger than previously reported when using more accurate fitness effect and evolutionary rate data [Wall DP, Hirsh AE, Fraser HB, in prep]), and here we have shown that they are also associated with a strategy of expression that maximizes the rate of transcription and minimizes the number of translations per mRNA. Given a desired rate of protein production, one way to maximize transcription rate while minimizing number of translations per mRNA is to maximize the mRNA decay rate. Thus, we would expect rapid mRNA decay among essential genes and protein complex subunits, which both evolve slowly, yielding the observed, negative correlation between the rates of mRNA decay and protein evolution. In support of this prediction, both essential genes and protein complex subunits have substantially shorter mRNA half-lives than the rest of the genome (e.g., nonessential genes have 32% longer mRNA half-lives than essential genes, and the bias remains when controlling for protein production rate; data not shown).

Discussion

We found that noise in protein production is minimized in genes for which it is likely to be most harmful, specifically essential genes and genes encoding protein complex subunits. This finding supports the hypothesis that noise in gene expression is generally deleterious to yeast.

Yeast appear to control the noise in their gene expression at both transcriptional and translational levels preferentially for some genes; however this noise minimization is not without a cost, as high transcription and high mRNA decay rates that are needed to minimize noise are energetically expensive, and are thus expected to be advantageous only when the benefit of reducing noise in a particular gene's expression outweighs this cost. Protein degradation rates may also play a role in controlling noise, but this cannot be tested until genome-wide protein degradation rates have been measured.

As is the case with many genome-wide studies, it is possible that a hidden variable could bias our results. For example, it is possible that essential genes and protein complex subunits tend to have high transcription and low translation for reasons unrelated to noise minimization. However, until such a variable can be identified, the most parsimonious interpretation of our results is that yeast adaptively minimize noise in the expression of certain genes.

As genome-wide transcription and translation rate data become available for other organisms, it will be interesting to see if the apparent tendency to minimize noise in the expression of important genes extends to organisms other than yeast. Considering that several anecdotal examples of indispensable genes with unusually low translation rates, and thus low noise in expression, have already been noted in *E. coli* (Ozbudak et al. 2002), this could well be the case.

Methods

Functional genomic data sources

Transcription rates (abbreviated as "txn rate") were calculated from mRNA abundances and decay rates in log-phase cells in rich glucose medium (Wang et al. 2002), according to the steady-state equation R=-ln(.5)*A/T, where R is transcription rate, A is mRNA abundance, and T is mRNA half-life. Translation rates per mRNA in rich glucose medium were calculated from ribosome occupancy data by Arava et al. (2003); specifically, ribosome density per mRNA present in the polysome fraction was multiplied by the fraction of each mRNA that was found in the polysome fraction, to estimate the average ribosome density for all copies of each mRNA in a cell. This density is equivalent to a relative translation rate, assuming that the speed at which ribosomes produces proteins is constant over different mRNAs. An estimate of the actual translation rate was found by multiplying the relative translation rates by a constant: the speed of translation, which is approximately 10 amino acids/sec (Arava et al. 2003). Protein production rate (proteins/sec) was then calculated by multiplying translation rate per mRNA with mRNA abundance. Note that the protein production rate can also be represented as the product of transcription rate and number of translations per mRNA (abbreviated as "tlns per mRNA"). It is this latter variable that was used to separate each bin into thirds in Fig. 2, since it is though to be more directly relevant to noise in protein production than related quantities such as translation rate per mRNA (Berg 1978); it was calculated by dividing protein production rate by transcription rate for each gene. However thirds could also be delineated by transcription rate, transcript abundance, or translation rate per mRNA, all yielding similar results (not shown).

Fitness effect ranks were calculated from 12 replicate growth experiments for all viable homozygous yeast deletion strains in rich glucose medium; growth experiments were conducted using the method described in Giaever et al. (2004). The logarithms of deletion strain tag fluorescence intensities on high-density oligonucelotide microarrays for each growth time course were fitted to a linear model that accounted for time course-specific effects and variable initial strain concentrations. The slope of the linear regression was then used as the relative growth rate for each strain.

Partial correlations

Partial correlations were calculated as described by Sokal and Rohlf (1995). Briefly, let r_{XY} be the correlation coefficient between variables X and Y. To control for a third variable Z,

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

To asses the significance of the partial correlation, it is transformed to be distributed according to a Student's t distribution, by the equation

$$t = r\sqrt{\frac{n-3}{1-r^2}}$$

The two-sided p-value can then be calculated according to where the t-value falls with respect to its expected distribution.

Footnote 1

Blake et al. (2003) showed that for two different promoters in yeast, as well as in their mathematical model, noise due to transcription peaked at ~1/3 of maximal transcriptional

induction. Importantly, one of their promoters (P_{ADH1*}) was 10-fold weaker than the other two at full induction, but all three showed very similar relationships between noise strength and % transcriptional induction. Since we do not have genome-wide data for the % induction for genes in rich glucose medium (or any other environment), in our analysis we make the assumption that the promoters of more highly transcribed genes tend to be at higher % induction levels. While this certainly does not hold for all genes, we believe that it is a reasonable approximation for most genes.

Footnote 2

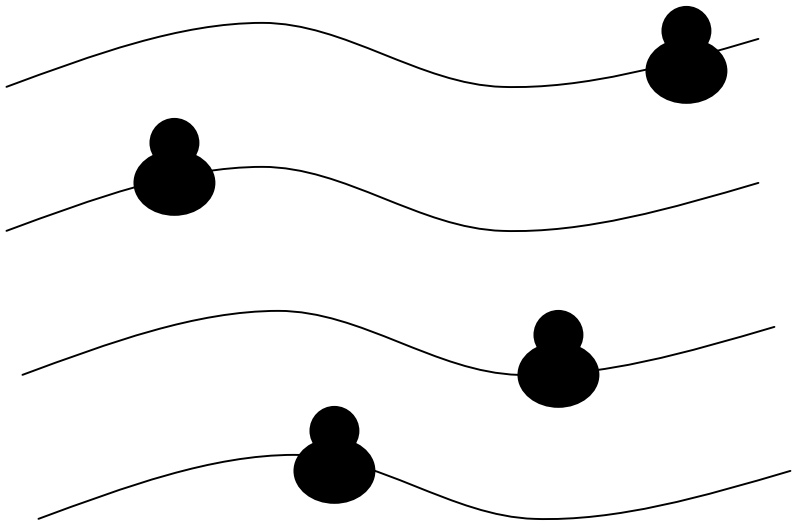
In the binning analysis the third of each bin with the lowest translation rate has, on average, a slightly lower overall protein synthesis rate than the third with the highest translation rate (not shown); this bias is the opposite of what would be expected from the association between protein synthesis rate and fitness effect or protein complex membership, and thus it acts against our observed bias to make the results of this analysis conservative underestimates of the true bias. In the case of protein complex subunits, it has been found that highly expressed genes are over-represented in protein complex data (whether this is an experimental artifact or a true relationship is unclear; von Mering et al. 2002); this would also act against our observed bias of complex subunits being over-represented in the third with the lowest overall protein synthesis rate in each bin, and thus make our results conservative.

Figure 1.

Strategies for expression. Three different strategies for achieving a given rate of protein production (four proteins will be produced in each case), and the amount of noise in which each strategy is expected to result. Curved lines represent mRNA molecules, with ribosomes translating them; more mRNA molecules represents higher transcription, and more ribosomes per mRNA represents higher translation per mRNA.

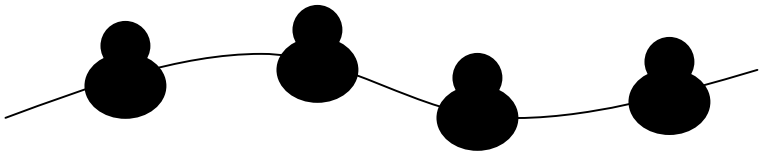
Strategy 1:

NOISE



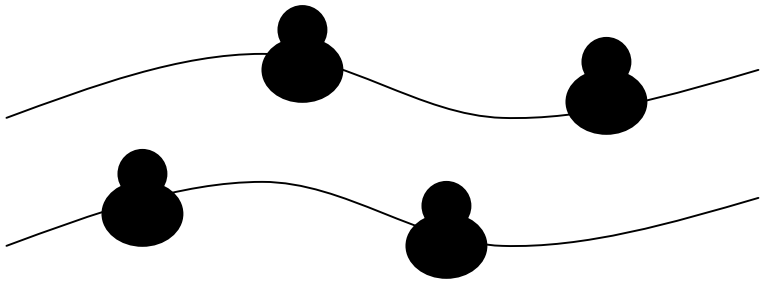
LOW

Strategy 2:



HIGH

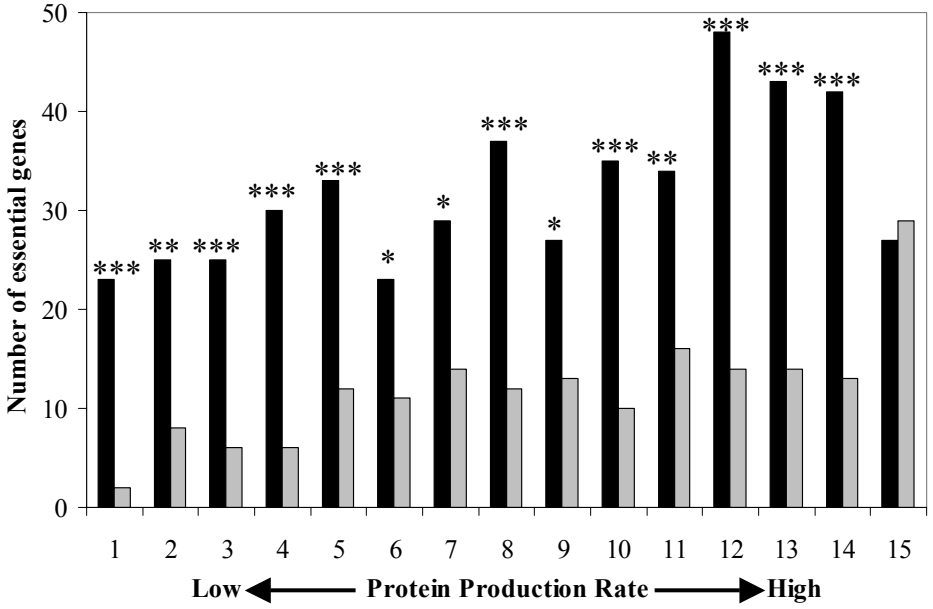
Strategy 3:

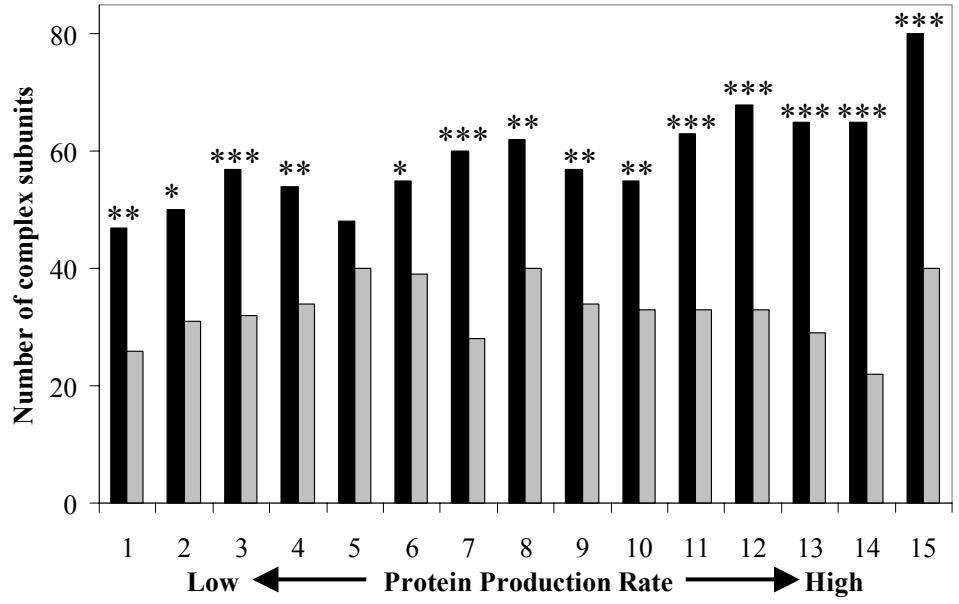


HIGH

Figure 2.

Essential genes and protein complex subunits minimize noise in expression. Binning analysis of (a) essential genes and (b) protein complex subunits. All genes for which transcription and translation rate data were available were separated into 15 bins by their protein production rate. Each bin was then separated into thirds by number of translations per mRNA. \ The two thirds in each bin with the most extreme transcription and translation are shown: black bars are the number of each type of gene (essential or complex subunit) in the third of each bin with the lowest number of translations per mRNA and the highest transcription rate, and thus low noise; gray bars are the number of each type of gene in the third with the highest number of translations per mRNA and the lowest transcripton rate, and thus high noise. Bins are ordered by their rate of protein synthesis. The number of asterisks indicates the Fisher's Exact Test probability of observing the values for each bin under the null model of independence. *, $P \leq 0.02$; **, $P < 0.005$; ***, $P < 0.0005$.


Number of essential genes
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Low
Protein Production Rate
High


Number of complex subunits
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Low
Protein Production Rate
High

Chapter VI

Evolution of aging in the primate brain.

The majority of this chapter is under review for publication, by:

Fraser HB, Khaitovich P, Plotkin JB, Paabo S, Eisen MB

Abstract

It is well established that gene expression levels in many organisms change during the aging process, and the advent of DNA microarrays has allowed genome-wide patterns of transcriptional changes associated with aging to be studied in both model organisms and various human tissues. Understanding the effects of aging on gene expression in the human brain is of particular interest, because of its relation to both normal and pathological neurodegeneration. Here we show that human cerebral cortex, human cerebellum and chimpanzee cortex each undergo different programs of age-related gene expression alterations. In humans, many more genes undergo consistent expression changes in the cortex than in the cerebellum; in chimpanzees, many genes change expression with age in cortex, but the pattern of expression changes bears almost no resemblance to that of human cortex. These results demonstrate the diversity of aging patterns present within the human brain, as well as how rapidly genome-wide patterns of aging can evolve between species; they may also have implications for the oxidative free radical theory of aging, and help to improve our understanding of human neurodegenerative diseases.

Introduction

Despite its ubiquity and importance, aging remains a poorly understood process. This lack of understanding is due in part to the complexity of aging, which is characterized by the gradual and progressive decline of numerous physiological processes and homeostasis (Pearl 1928; Harman 1956; Beckman and Ames 1998; Hekimi and Guarente 2003; Poon et al 2004; Stadtman 2001). However, recent progress in aging

research has made it clear that aging processes are amenable to biochemical and genetic dissection, in both humans and model organisms (Beckman and Ames 1998; Hekimi and Guarente 2003; Poon et al 2004; Stadtman 2001).

The most widely held mechanistic theory of aging is known as the free radical theory, and was first introduced by Harman almost half a century ago (Harman 1956). This theory, as well as the related "rate of living" theory proposed earlier by Pearl (1928), holds that aging is due to deleterious side effects of aerobic respiration. Specifically, mitochondrial activity leads to the production of reactive oxygen species (ROS) that can damage many cellular components, including DNA, lipids, and proteins (Beckman and Ames 1998). These ROS, such as the hydroxyl radical (OH•) and hydrogen peroxide (H_2O_2), are produced in large part by the mitochondrial electron transport chain. The free radical theory has garnered widespread support in recent years; in particular, studies from a number of model organisms showing that decreasing ROS levels leads to an increase in lifespan indicate that ROS are a cause, and not just a correlate, of aging (Beckman and Ames 1998; Hekimi and Guarente 2003). As further evidence in favor of the theory, it is known that decreasing metabolic rate through caloric restriction extends lifespan in many organisms (Beckman and Ames 1998; Hekimi and Guarente 2003).

Exactly how macromolecules damaged by ROS may lead to aging has been studied in detail in recent years, and the human brain has been intensively examined in this regard because of its overall importance in human senescence. Up to one-third of the proteins in the brains of elderly individuals may be oxidatively damaged, and these damaged proteins have been shown to sometimes have diminished catalytic function (Beckman and Ames 1998; Stadtman 2001). One recent study of aging in the human

brain demonstrated that oxidative damage to DNA can be caused by mitochondrial dysfunction, and tends to accumulate preferentially in some areas of the genome that include promoters, resulting in lower levels of transcription (Lu et al 2004) (possibly due to loss of transcription factor or other protein binding [Ghosh et al 1999; Marietta et al 2002; Brooks et al 2000]). In this same study, genome-wide patterns of aging-associated gene expression change in one region of the human brain cortex (the frontal pole; Figure 1) were measured using DNA microarrays, and genes that had decreased transcription with age were shown to be the ones that are most susceptible to oxidative damage (Lu et al 2004). Since different regions of the human brain have been shown to accumulate DNA damage at different rates (Corrall-Debrinski et al 1992; Mecocci et al 1993), it is reasonable to suppose that these different regions may show different gene expression changes with age as a result.

Complementing studies of aging in single species, research into the evolution of aging has been helpful in advancing our understanding of senescence (Partridege and Barton 1993). Aging and lifespan can in principle be subject to natural selection, since artificial selection on model organisms in the lab can lead to dramatic changes in lifespan in a very small number of generations (Partridege and Barton 1993). Furthermore, some of the proteins involved in aging seem to play an evolutionarily conserved role; for example, the histone deacetylase SIR2 has been shown to affect aging in organisms as diverse as yeast, nematode, fruit fly, and mouse (Hekimi and Guarente 2003). However while the factors that cause aging (such as ROS) and control the rate of aging (such as SIR2) appear to be highly conserved, it is not as clear whether the consequences of aging are equally conserved on a molecular level.

One promising approach to answering this question lies at the level of gene expression: do orthologous genes tend to undergo the same patterns of expression changes with age in diverse species, or can a common factor such as ROS lead to different gene expression patterns in different organisms? Using DNA microarrays, this question can now be addressed in a systematic, genome-wide manner. One notable study found that a small but significant portion of aging-related gene expression changes are shared by the very distantly related nematode and fruit fly (McCarroll et al 2004); another study comparing aging patterns in muscle cells of two more closely related species, mouse and human, also found a great deal of divergence in aging patterns (Welle et al 2001). While both of these studies are informative, neither address the questions of how quickly age-related gene expression patterns can evolve over short periods of time, and if humans in particular show unique patterns of aging not shared by closely related primates.

The human brain is of particular interest for studying the divergence in phenotypes that have changed rapidly during evolution (such as aging). Brain-specific genes have undergone accelerated evolution in the lineage leading to human since the split with chimpanzee at the levels of both protein sequence and gene expression (Enard et al 2002; Dorus et al 2004), pointing to the numerous functional differences that have accumulated between these two species since their divergence only five to seven million years ago. Aging in the human brain is also of interest because ROS-induced damage and age are both major risk factors in many neurodegenerative diseases (such as Alzheimer's, Parkinson's, and Amyotrophic Lateral Sclerosis [Emerit et al 2003]). In this study we addressed two questions about the relationship between gene expression

and aging. First, using published data, we asked whether the pattern of gene expression change with age previously observed in the frontal pole (Lu et al 2004) is representative of other regions of the human brain. Then, using data generated for this project, we asked how similar the senescence-associated changes in gene expression observed in human brain (Lu et al 2004) are to those observed in our closest living relative, the chimpanzee.

Results

In order to test whether different regions of the human brain show similar patterns of change with age, we utilized three independent published microarray expression data sets. These were: Lu et al (2004), mentioned above, in which the frontal pole regions of 30 individuals (aged 26-106 years) were used to identify hundreds of genes with clear up- or down-regulation associated with age; Khaitovich et al (2004), in which gene expression patterns of six brain regions (Figure 1: prefrontal cortex, primary visual cortex, anterior cingulate cortex, Broca's area, caudate nucleus, and cerebellum) were studied in three individuals (aged 45, 45, and 70 years); and Evans et al (2003), in which three brain regions (Figure 1: prefrontal cortex, anterior cingulate cortex, and cerebellum) from seven individuals (aged 18-70) were studied. The latter two studies were conducted to examine gene expression differences between regions of human brain; the data were not previously analyzed with respect to aging. All three studies used the same microarray platform (Affymetrix HG U95Av2), facilitating comparison between them.

Aging is heterogeneous within the human brain

To achieve the most comprehensive picture of brain aging possible with these data, we first sought to study the patterns of aging in all six brain regions from Khaitovich et al (2004). Since only three samples of two ages were available for each brain region in this data set, only three general aging patterns were possible: up-regulation (the old sample is more highly expressed than either young sample), down-regulation (the old sample is more weakly expressed), or neither (the old sample is in between the young samples). Because thousands of genes would be expected to show each of these three patterns even in the absence of any genuine aging-related changes in gene expression, we were unable to use the three samples on their own to accurately identify genes changing expression with age.

However, with the available data we could ask whether the genes whose expression changes with age in frontal pole (Lu et al 2004) showed the same direction of change in each of six other brain regions (Khaitovich et al 2004). In order to do this, we reanalyzed the data of Lu et al (Lu et al 2004), and identified 841 genes that showed a significant (p<0.01) Spearman rank correlation between age and expression level in frontal pole; most of these were expected to be true positives, since only ~126 genes would be expected to pass this significance threshold by chance. We classified these 841 genes as having either increasing or decreasing expression with age in frontal pole, and then as either increasing, decreasing, or constant in each of the six other brain regions. After discarding genes with no direction of change within each of the six brain regions, since these lack any information about aging changes, we tested how well the frontal pole data agree with the data from each of the six other regions. For example, comparing

prefrontal cortex to frontal pole, we asked how many genes belong to each of four categories: 1. up-regulated in frontal pole and down-regulated in prefrontal cortex; 2. down-regulated in frontal pole and up-regulated in prefrontal cortex; 3. up-regulated in both regions; 4. down-regulated in both regions. If the data sets showed similar aging patterns, we would expect an excess of genes in the latter two categories, whereas no such excess would be expected in the absence of a shared pattern. There are a number of statistical tests that can be used to quantify these patterns; we chose to use the nonparametric Spearman rank correlation coefficient (abbreviated as r). Values of r close to one indicate good agreement between aging patterns, whereas those close to zero indicate a lack of agreement. To assess the significance of these correlations, we randomly permuted the ages of the samples, and calculated the probability of observing a random correlation as strong as that found in the real data (see Materials and Methods).

Strikingly, all four regions of cerebral cortex for which we had expression data (prefrontal cortex, Broca's area, primary visual cortex, and anterior cingulate cortex) showed excellent agreement with the aging pattern from frontal pole (Figure 2A; $r>0.8$ and $p<0.02$ for each). In sharp contrast, cerebellum and caudate nucleus showed far less agreement with frontal pole (Figure 2A; $|r|<0.1$ and $p>0.4$ for each). These results have several implications. First, the agreement between frontal pole and four regions of cortex indicates that we are able to accurately measure the direction of gene expression changes for most genes, even with only three samples from each region; thus the age range, number of samples, etc., are all sufficient to reflect the pattern of gene expression changes previously reported in frontal pole (Lu et al 2004). Second, we can have even greater confidence in the results from frontal pole (Lu et al 2004), since they have been

independently reproduced (albeit in different brain regions). Third, and most importantly, the human brain appears to have different aging patterns in cerebellum and caudate nucleus than in cortex. The fact that our four cortex samples all show strong correlations with frontal pole is akin to having a positive control, and it allows us to interpret the lack of correlation in cerebellum and caudate nucleus as evidence suggesting a difference in aging patterns, as opposed to several more trivial explanations (e.g., too few samples).

In order to further test the similarity of aging patterns within the brain, we compared a third independent data set to the data from Lu et al (2004) and Khaitovich et al (2004). As described above, Evans et al (2003) sampled three brain regions from each of seven individuals. We first tested whether the aging patterns in the two cortex regions from Evans et al (2003) correlated more highly with the frontal pole aging changes (Lu et al 2004) than did the cerebellum samples, as would be expected from Figure 2A. Classifying the same 841 genes showing significant change with age in the frontal pole as either up-regulated or down-regulated with age in each brain region of this new data set, we found the same general pattern of correlations as with the data from Khaitovich et al (2004): cerebellum showed a weaker correlation with frontal pole than did either cortex sample (prefrontal cortex, r=0.70; anterior cingulate cortex, r=0.61; cerebellum, r=0.38). While the cerebellum correlation is stronger here than in Figure 2A, it is still not significantly different from zero (p=0.17), though the two cortex samples are both significant (p<0.01 each). This finding supports our conclusion that cerebellum ages differently than cortex.

For a third test of aging patterns throughout the human brain, we determined the correlation of aging patterns between a single cortex region from Evans et al (2003) with all six of the brain regions from Khaitovich et al (2004). We used the prefrontal cortex samples from Evans et al (2003), since as mentioned above, this brain area shows a better agreement of aging patterns with frontal pole than does anterior cingulate cortex. To facilitate comparison with cerebellum (see below), we extended this analysis to all 12,558 probe sets present on the microarray; however in order to increase the signal/noise ratio, we then excluded genes with no apparent aging changes (age vs. expression $|r|<0.5$) in the Evans et al (2003) data. This comparison showed the expected reproducibility of aging patterns across all four regions of the cortex: $r>0.76$ for all four (Figure 2B; $p<0.03$ for each except for PFC, for which $p=0.067$). In contrast, neither cerebellum nor caudate nucleus showed a significant correlation (Figure 2B; $|r|<0.04$ and $p>0.4$ for each), as expected from their lack of correlation with the frontal pole data shown in Figure 2A. In addition to providing further support for our finding of an aging pattern common to all tested regions of cortex, this result demonstrated that even when comparing aging patterns from the two smaller microarray studies used here (Khaitovich et al 2004; Evans et al 2003), the age range, number of samples, etc., were sufficient to reveal a correlation if one exists.

The lack of correlation between the aging pattern in cerebral cortex with those in cerebellum and caudate nucleus might arise because the quality of data in the cerebellum and caudate nucleus samples was lower than that of the cortex samples in both Khaitovich et al (2004) and Evans et al (2003); lower quality of data would lead to weaker correlations. To address this possibility, we first compared the expression levels

of the 841 genes used in Figure 2A in the two 45 year-olds from Khaitovich et al (2004), since their equal age controls for the fact that we expect these genes not to have a very high correlation between sample of different ages (such as between the 45 and 70 year-olds). All six brain regions had highly reproducible expression levels; the lowest correlation among all six was for anterior cingulate cortex, with r=0.952. The cerebellum data from Evans et al (2003) was of similarly high quality: among five replicates of the same cerebellum samples analyzed in two different laboratories, the lowest correlation of expression levels among all genes was r=0.964. Thus differing data quality could not explain the lack of correlation in cerebellum and caudate nucleus.

Human cerebellum ages less than cortex

There are two possible explanations for the difference in the aging patterns between cerebellum/caudate nucleus and cerebral cortex. One is that cerebellum and caudate nucleus have their own aging patterns distinct from that in cortex. The other possibility is that cerebellum and caudate nucleus are different from cortex because they each have far *fewer* genes changing expression with age than cortex does, and they thus lack a reproducible pattern of aging-associated gene expression changes altogether.

To distinguish between these possibilities, one could attempt to calculate exactly how many genes change expression with age in each region; if cerebellum and/or caudate nucleus have aging-related changes in as many (but a different set of) genes than cortex, then the number of genes identified as changing in cerebellum and/or caudate nucleus should be comparable to any region of cortex. Unfortunately, as mentioned above, there

is not enough statistical power to pursue this approach, given only three samples per region (or seven, as in Evans et al [2003]).

Another way to differentiate between the two possibilities listed above would be to compare two data sets of cerebellum and/or caudate nucleus aging patterns to one another. If these regions have a reproducible pattern of many genes changing expression with age (as in the cortex samples of Figure 2A-B), we should find a significant correlation. A comparison between the data from Evans et al (2003) and Khaitovich et al (2004) is suitable for this purpose, since both data sets contain cerebellum samples, and since we already have a positive control that demonstrated our ability to find a correlation between aging patterns in these data sets when one exists (Figure 2B).

We thus expected to see a strong positive correlation between cerebellum aging patterns in our two data sets if and only if a large number of genes change expression with age in cerebellum. Because this analysis was carried out on all informative genes (age vs. expression $|r|>0.5$, as in Figure 2B), instead of just the 841 with expression changes in the frontal pole, any reproducible changes in cerebellum should be found. Comparison of cerebellum aging from Evans et al (2003) with all six regions from Khaitovich et al (2004) gave an unambiguous result: not a single region had a significant correlation (Figure 2C; $|r|<0.2$ and $p>0.4$ for each), including the cerebellum-cerebellum comparison. From these results, we conclude that cerebellum has a different pattern of aging than cortex because significantly fewer genes appear to change expression with age in cerebellum.

Chimpanzee cortex ages differently than human cortex

In order to further characterize the differences in aging patterns between cortex and cerebellum, we calculated the average expression levels in one representative region of cortex (prefrontal cortex) for the 841 genes changing strongly with age in frontal pole, in both young (45 year old) and old (70 year old) samples from Khaitovich et al (2004). As expected, when separated into two groups by their direction of change with age, clear differences were seen between the young and old samples (Figure 3, red lines). When the same genes were subjected to this analysis using their cerebellum expression levels, an interesting trend emerged: while the genes up-regulated in cortex (Figure 3, red dashed line; Wilcoxon p=0.028 comparing young vs. old) are also slightly up-regulated in cerebellum (Figure 3, blue dashed line; Wilcoxon p=0.062), those that are down-regulated in cortex show no significant change at all in cerebellum (Figure 3, blue solid line; Wilcoxon p=0.57). Thus, the difference in aging patterns between these two brain regions arises mainly from genes down-regulated in the cortex. The reason for this may be related to metabolic differences between cortex and cerebellum (see Discussion).

In order to study the relationship between brain aging patterns in humans and chimpanzees, we required gene expression data from the chimpanzee brain. While four studies have already produced such data (Enard et al 2002; Khaitovich et al 2004; Uddin et al 2004; Caceres et al 2003), three of these examined only a single brain area, and the fourth had an insufficient number of samples of appropriate age for our purposes. Therefore, we generated new data by measuring gene expression levels in three regions of the chimpanzee brain: prefrontal cortex, anterior cingulate cortex, and cerebellum (see Materials and Methods). For each region we had five samples from the same individuals

(aged 7 to ~45 years; see Materials and Methods), as well as one additional sample for cerebellum and two for prefrontal cortex (though excluding the three extra samples made little difference in the analysis; see Materials and Methods). Because of the very high sequence similarity between humans and chimpanzees (Sakaki et al 2003), we were able to use microarrays designed for human sequences; since we are only comparing chimpanzee samples directly with one another (comparisons with human are using only aging patterns, not actual expression levels), masking of microarray probes containing DNA sequence differences between human and chimpanzee was not necessary (and did not affect the analysis when tested).

We compared the aging patterns of human and chimpanzee brains by applying the same methods as for comparison between aging patterns of different regions within the human brain. Using the 841 genes that change expression with age in frontal pole (Lu et al 2004), we tested the agreement between directions of aging changes in frontal pole and changes in each of our three chimpanzee brain regions. As can be seen in Figure 4A, none of the three regions showed any significant correlations with human frontal pole ($|r|<0.13$, $p>0.4$ for all three). Similar results were found when comparing chimpanzee aging in any brain region to the patterns from either of our other two human expression data sets [Khaitovich et al 2004; Evans et al 2003], for either the 841 genes or all genes on the microarray (not shown). While these results are consistent with chimpanzees having very different aging patterns than humans, many other possibilities could not be ruled out without more evidence. For example, the chimpanzee data could be less accurate than human data due to the microarray being designed for human, or the age range of our chimpanzee samples may not be great enough to reflect aging changes.

To distinguish between these alternatives, we compared the chimpanzee aging patterns in the three different brain regions directly with one another. As in Figure 2B-C, we used all genes present on the microarray, except for uninformative genes (age vs. expression $|r|<0.5$). While we found no significant correlation when comparing cerebellum to either cortex region (Figure 4B; $|r|<0.07$ and $p>0.3$ for both comparisons), similar to the case for human shown in Figure 2, we found a very strong agreement when comparing aging patterns of prefrontal cortex to anterior cingulate cortex (Figure 4B; $r=0.894$, $p<0.005$). This is a crucial result, since it eliminates virtually all possible explanations save one: chimpanzee cortex has a reproducible pattern of aging-associated gene expression changes, but this pattern is completely different from that of human cortex. Alternative explanations such as lower chimpanzee data accuracy, insufficient chimpanzee age range, and very few age-related changes in chimpanzee cortex (as in human cerebellum) can all be eliminated, since they would all preclude a strong correlation between aging patterns of the two chimpanzee cortex regions.

Given this difference in aging patterns between humans and chimpanzees, we examined the expression levels of the chimpanzee orthologs of the 841 human genes that change expression with age in frontal pole in order to see if the expression levels of the chimpanzee orthologs of these genes resemble young humans, old humans, or neither. To test this, we first reanalyzed the expression data by masking all microarray probes with sequence differences between humans and chimpanzees (Khaitovich et al 2004). We then calculated, for both human and chimpanzee prefrontal cortex, the average expression level for the set of genes that increase expression with age in frontal pole, as well as the average for the genes that decrease expression with age. The result is that in

chimpanzee cortex, the orthologs of both sets of human genes (up-regulated and down-regulated) are expressed at the levels of their young human counterparts (Figure 5). In other words, chimpanzee cortex expression levels strongly resemble expression levels in young but not old humans, at least among the set of genes tested here. Humans then diverge from these average expression levels as they age, whereas chimpanzee gene expression levels change in an almost entirely different set of genes.

The high correlation of gene expression aging patterns between the two regions of chimpanzee cortex implied that the genes for which both regions show the same direction of change (that passed our cutoff of age vs. expression $|r|>0.5$) are nearly all genuinely up- or down-regulated with age. Using this list of genes with a consistent aging pattern in the two cortex regions, there were 1252 down-regulated and 700 up-regulated genes. Note that while the false-positive rate is likely to be low, we have no way to estimate the false-negative rate, so these numbers should not be interpreted as the total number of genes changing expression with age in chimpanzee cortex. Using this list of aging-associated genes, we tested for any significant enrichments of these genes in Gene Ontology annotation categories (Khaitovich et al 2004). We did not find any enrichments for the set of genes down-regulated with age, though we found a number of significant enrichments for those up-regulated with age, including mitochondrial localization, protein degradation functions, and several metabolic processes (see Supplementary Table 1). Interestingly, and consistent with our finding of no similarity between human and chimpanzee aging patterns, there was little overlap between these enriched groups and those previously reported for human frontal pole (Lu et al 2004).

Discussion

In this study, we have made three main observations: first, aging-related gene expression changes are similar throughout all five tested regions of the human cerebral cortex. Second, human aging changes are quite different in cerebellum and caudate nucleus, and at least in cerebellum this appears to be due to far fewer genes changing expression with age than in cortex. Third, while chimpanzee cortex has a reproducible pattern of expression changes with age, it shares no detectable similarity with the aging pattern in human cortex.

These conclusions raise a number of questions. For example, why does human cerebellum age differently than human cortex? We have shown that the majority of the difference is because genes down-regulated with age in cortex are not down-regulated in cerebellum (Figure 3); since fewer genes change with age in cerebellum than in cortex (Figure 2), it therefore follows that this is mostly due to fewer genes being down-regulated with age. What could cause fewer genes to have reduced transcription over time in cerebellum than in cerebral cortex? Cerebellum differs in many important respects from cortex; in particular, it has a lower metabolic rate than cortex in both human and rhesus macaque, regardless of age (Sakamoto et al 1999; Bentourkia et al 2000; Noda et al 2002). These observations of lower metabolic activity in cerebellum imply that if a consequence of aerobic respiration is ROS-induced DNA damage, then such damage should be greater in cortex than in cerebellum. Indeed, it has been shown that cerebellum has far fewer mitochondrial DNA (mtDNA) deletions than cortex, especially in old humans (Corral-Debrinski et al 1992), and it accumulates less oxidative damage to both mtDNA and nuclear DNA than does cortex (Mecocci et al 1993).

Therefore if the accumulation of DNA damage causes gene expression down-regulation, then we would expect to see fewer aging-related gene expression reductions in cerebellum than in cortex. We interpret our confirmation of this prediction as evidence in support of the theory that ROS-induced damage is responsible for gene expression changes (Lu et al 2004; Evans and Cooke 2004), as well as the more general oxidative free radical/"rate of living" theory of aging (Pearl 1928; Harman 1956).

Similarly, one might ask why chimpanzee cortex ages differently than human cortex. If we assume that ROS-induced DNA damage is a major cause of gene expression changes (Lu et al 2004; Evans and Cooke 2004), then the question becomes, why are the areas of the chimpanzee genome damaged by ROS different than in human (while ROS damage is unlikely to directly explain the difference in up-regulated genes seen in Figure 5, it may be indirectly responsible, by down-regulating genes such as transcriptional repressors)? This question is difficult to answer, since we do not presently understand what factors lead to ROS damage susceptibility; regardless of the factors involved, however, it is quite possible that even the relatively few differences in DNA sequence between human and chimpanzee may be sufficient to cause drastic changes in ROS susceptibility, as is the case for other chromosomal properties such as DNA methylation (Enard et al 2004) and recombination rate (Ptak et al 2004; Winckler et al 2005). One possible explanation for the divergence is that promoters driving high levels of transcription are more susceptible, perhaps because of their more accessible chromatin structure and/or lower tolerance for oxidative damage; however while highly expressed genes are indeed more likely to be down-regulated with age in human frontal pole, expression levels are far from explaining all of the variation in aging-related changes in

either human or chimpanzee (not shown). It may also be that other factors such as subnuclear localization of different chromosomal regions is involved in damage susceptibility and that this differs between humans and chimpanzees; since different mammalian cell types have distinct nuclear organizations of chromosomes (Parada et al 2004), it will be interesting to see if ROS damage susceptibility in different cell types can be correlated with the nuclear organization of chromosomes in those cells.

Another implication of these results is related to the use of model organisms such as mouse, rat, and various primates as surrogates for human brain aging and neurodegeneration. The fact that even the chimpanzee, our closest living relative, has patterns of age-related gene expression changes almost entirely different than human implies that making specific inferences about human brain aging from model organisms may be difficult. This is supported by a study of brain aging in mice, where in contrast to the results reported here for human, mouse cerebellum was found to contain *more* genes changing expression with age than mouse cortex (Lee et al 2000), a difference that may be due to different relative metabolic rates of cortex and cerebellum in mouse compared to human. Model organisms are probably well-suited for studying the mechanisms of aging (such as ROS-induced damage), which are likely to be conserved over great phylogenetic distances, but such conserved mechanisms may have species-specific outcomes at the level of individual genes. Thus, caution is warranted when trying to extrapolate the results of neurodegeneration research from model organisms to humans.

Many other questions raised by this work are still unresolved. First, how diverse are aging patterns of gene expression change in human tissues outside the brain? A recent study finding similar aging profiles of human kidney cortex and medulla regions

implies that the intra-organ variability in aging patterns observed in the present work may not be found in all organs (Rodwell et al 2004). Second, do human and chimpanzee differ in their aging patterns in tissues other than brain, or is the brain a special case because of its recent rapid morphological evolution in the human lineage? It will be interesting to test this for tissues which have not undergone any obvious rapid evolution (such as liver or kidney), as well as for tissues which are likely to have been under strong positive selection (such as testes). Third, does chimpanzee cerebellum have fewer gene expression changes with age than cortex, as is the case in human? More chimpanzee data will be needed to address this question, though it seems likely that the answer will be affirmative, since the greater metabolic rate of cortex compared to cerebellum is conserved to rhesus macaque (Noda et al 2002). Fourth, does the human or chimpanzee cortex aging pattern represent the ancestral state of this pattern for these two species, or are they both highly diverged from that state? Examination of brain aging patterns in an outgroup species, such as rhesus macaque, may help to resolve this question. Fifth, is the rapid divergence of aging patterns along the human and/or chimpanzee lineage the result of selection on the aging process itself, or is the divergence an indirect consequence of selection on other aspects of the brain, or could it even be explained by random drift alone? And finally, can we use our understanding of the similarities and differences in brain aging of humans and chimpanzees to gain greater insight into the causes of, and possible treatments for, human neurodegeneration? We believe this will be possible, since investigation of how a phenomenon such as neurodegeneration emerged during evolution might well point us towards its underlying causes.

Materials and Methods

Tissue samples and gene expression data

Human microarray data was obtained from Pritzker Neuropsychiatric Disorders Research Consortium (http://www.pritzkerneuropsych.org) and the NIMH Silvio O. Conte Center, and from ArrayExpress and GEO databases. Only the seven "Type 1" control individuals (Li et al 2004; Tomita et al 2004) were used from the Evans et al (2003) dataset.

Chimpanzee postmortem samples were obtained from Yerkes Regional Primate Center, Biomedical Primate Research Centre and Anthropologisches Institut und Museum, Universität Zürich. All individuals suffered sudden death for reasons other than their participation in this study and without any relation to the tissues used. Total RNA was isolated from approximately 50 mg of frozen tissue using the TRIZol® reagent according to manufacturer's instructions and purified with QIAGEN® RNeasy® kit following the "RNA cleanup" protocol. RNAs were of high and comparable quality in all samples as gauged by the ration of 28S to 18S ribosomal RNAs estimated using the Agilent® 2100 Bionalyzer® system and by the signal ratios between the probes for the 3' and 5' ends of the mRNAs of GAPDH and β-actin genes used as quality controls on Affymetrix® microarrays. Labeling of 1.2 microgram of total RNA, hybridization to Affymetrix® HG U95v2 arrays, staining, washing and array scanning were carried out following Affymetrix® protocols. The samples were processed in random order with respect to age. All primary expression data are publicly available at ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/). Data were normalized using the Robust Multichip Average (RMA) method (Bolstad et al 2003).

Among the chimpanzees used for this work, one was of indeterminate age, having been caught in the wild 40 years before its death. However since we used nonparametric rank statistics for all analyses, the exact age was irrelevant; all that mattered was whether it was older or younger than our 44 year-old chimpanzee. For the results shown it was assumed to be older, though conducting the analyses assuming it to be younger than 44 years strengthened the results (the correlation of chimpanzee PFC-ACC increased from 0.894 to 0.927).

Additionally, one pair of chimpanzees used were full siblings, and another pair were half siblings, which could be problematic if aging patterns are family-specific. However one member of each related pair could be excluded from the analysis without greatly affecting the results (correlation of chimpanzee PFC-ACC decreased from 0.894 to 0.858), indicating that relatedness among chimpanzees does not affect our results. Excluding one member of each related pair also left the same five unrelated chimpanzees for each of the three brain regions tested, controlling for any possible effects of unequal sample sizes from each brain region.

There are several possible artifactual explanations for our results not addressed in the main text. First, there is the possibility that gene expression changes correlated with age were caused by an unknown factor unrelated to the normal aging process. For example, if in the study by Khaitovich et al (2004) the 70 year-old had a disease which made his cerebellum and caudate nucleus appear "young" in their gene expression, but did not affect his cortex (since all four of his cortex regions showed the same reproducible pattern), then this could account for the results of Figure 2A. However in order for this explanation to also account for the similar results of Figure 2B, several

elderly individuals from the Evans et al (2003) study would all have to be similarly afflicted in their cerebella as well. We found this to be extremely unlikely, since all ten individuals from these two studies were chosen in part for their lack of any known brain-related diseases (Khaitovich et al 2004; Evans et al 2003).

Similarly, one possible explanation for the results in Figure 4B is that the cortexes (but not cerebella) of both of our old chimpanzees had a large number of gene expression differences compared to the young chimpanzees for a reason other than aging, such as a cortex-specific brain disease distinct from the normal aging process. As for the human subjects discussed above, this is extremely unlikely to be the case, since none of these chimpanzees had any apparent brain disease, and *both* old chimpanzees (who are unrelated to each other) would have to be similarly afflicted to observe this effect.

Another possible explanation for the difference between human and chimpanzee aging patterns (Figure 4A) is that the difference is actually due to the different environments experienced by the humans and chimpanzees during their lifetimes, and is not due to any intrinsic differences between these species. Because there is no way to possibly control for this, since for both practical and ethical reasons a human cannot be raised in precisely the same environment as a chimpanzee (or even another human), all that we can rigorously conclude is that the humans and chimpanzees used for the analyses herein did experience different patterns of gene expression change with age. We note that this general concern extends to all studies comparing any human's phenotype with that of another organism.

One caveat concerning our interpretation of fewer genes having aging-associated changes in expression in cerebellum than in cortex is that the two human cerebellum data

sets, while both consisting of grey matter of the cerebellum, were from different regions of the cerebellum (Khaitovich et al [2004] sampled the Vermis cerebelli, whereas Evans et al [2003] used the left lateral portion of the cerebellum); therefore it is technically possible that these regions of cerebellum each have their own reproducible aging pattern, which (if both shared no similarity whatsoever either to each other or to cortex) would not be revealed by this analysis. We find this to be quite unlikely, given the very close proximity and functionally similar properties of these two regions, together with the finding that far more heterogeneous regions throughout cortex share nearly identical aging patterns. And even if this improbable case were to be true, our conclusion of cerebellum grey matter *as a whole* lacking any reproducible aging pattern would still hold.

Statistics

All correlation coefficients reported here were calculated by Spearman's rank correlation, a nonparametric method which is robust to the presence of any outliers. The correlation coefficients from comparisons of aging profiles between two tissues or species (as in Figures 2 and 4) can be interpreted as scores directly proportional to the fraction of genes with the same direction of expression change with age. Probability values were calculated by randomization of ages: the fraction of randomizations with a correlation coefficients greater than or equal to the observed value is the *p*-value given. Therefore this is a one-sided test, appropriate for the question of whether we could have agreement between aging patterns as strong as those observed, just by random chance. The only exceptions to our using this one-sided test were when we stated we were testing

whether the correlation coefficient was significantly different from zero (as opposed to greater than zero); in these cases the test was two-sided. We note that this randomization test is very conservative; for example, analyzing the leftmost bar of Figure 2A (PFC) with Fisher's exact test yields a *p*-value of $\sim 10^{-99}$, as opposed to ~0.015 from randomization. This large difference is due to the nonrandom structure of the expression data, which makes it more likely to observe strong correlations than would be expected in a set of random data.

Figure 1.

The seven regions of the human brain analyzed in this work. Abbreviations used in other figures are: FP, frontal pole; PFC, prefrontal cortex; PVC, primary visual cortex; ACC, anterior cingulate cortex; BA, Broca's area; CN, caudate nucleus; C, cerebellum.

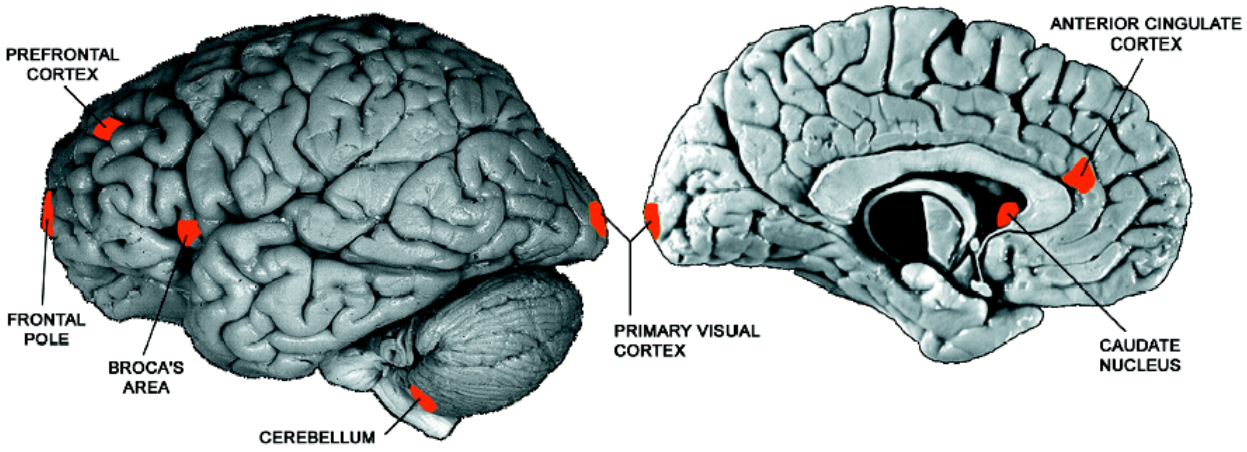

PREFRONTAL CORTEX
FRONTAL POLE
BROCA'S AREA
CEREBELLUM
PRIMARY VISUAL CORTEX
ANTERIOR CINGULATE CORTEX
CAUDATE NUCLEUS

Figure 2.

Aging in the human brain. (A) Correlations of aging gene expression patterns between human frontal pole (Lu et al 2004) and each of the six regions of the human brain from (Khaitovich et al 2004). The high correlation for all four cerebral cortex samples indicates a reproducible aging pattern across all tested regions of cortex; this pattern does not hold for caudate nucleus or cerebellum. (B) Correlations of aging gene expression patterns between human prefrontal cortex (Evans et al 2003) and each of the six regions of the human brain from (Khaitovich et al 2004). The high correlations for all four cortex samples indicates a reproducible aging pattern across all tested regions of cortex but not caudate nucleus or cerebellum, confirming the result of Figure 2A. (C) Correlations of aging gene expression patterns between cerebellum (Evans et al 2003) and each of the six regions of the brain from (Khaitovich et al 2004). The lack of any significant correlation, even when comparing the two cerebellum aging patterns to each other, indicates that human cerebellum lacks a reproducible aging pattern.

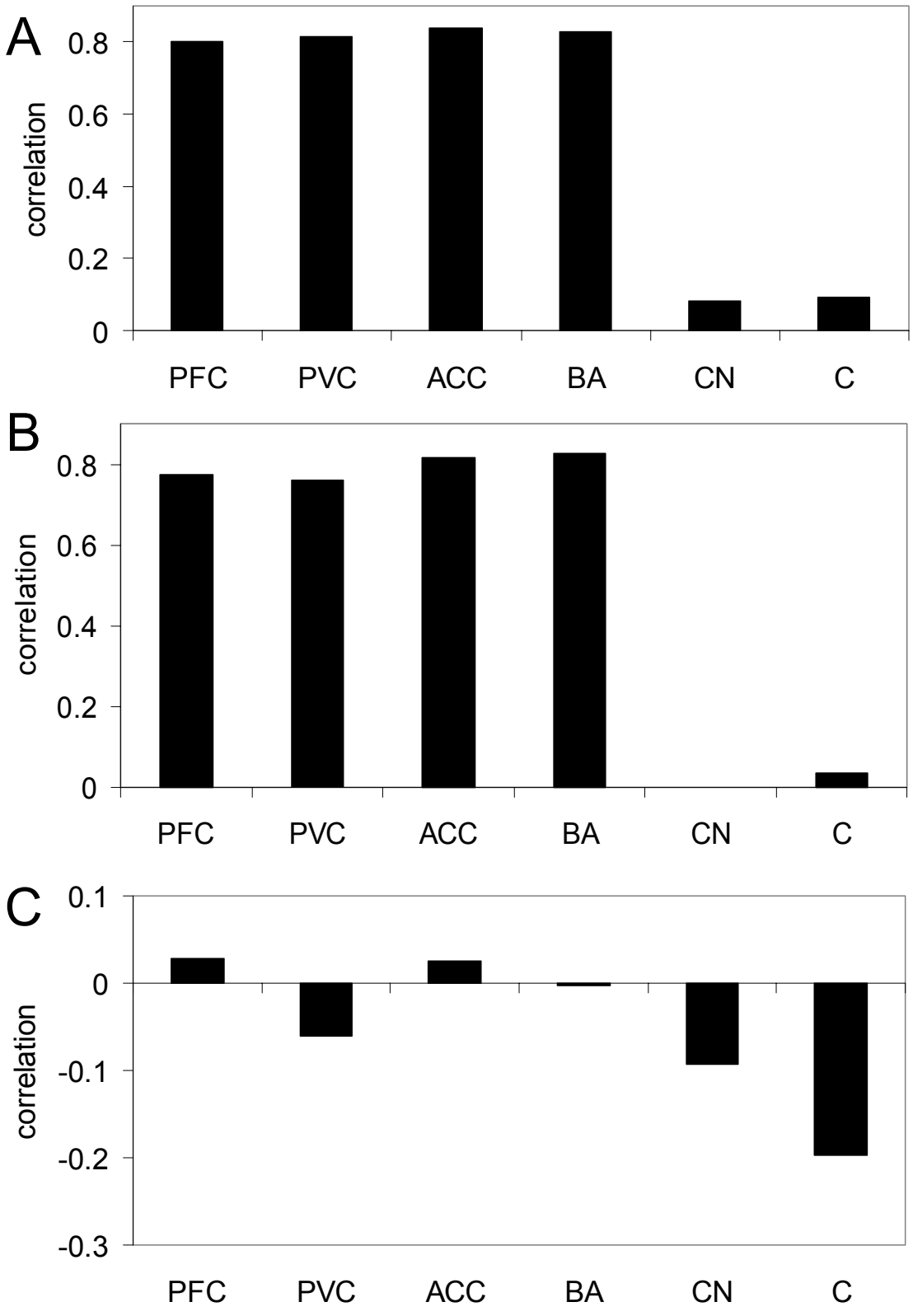

A
correlation
0.8
0.6
0.4
0.2
0
PFC
PVC
ACC
BA
CN
C
B
correlation
0.8
0.6
0.4
0.2
0
PFC
PVC
ACC
BA
CN
C
C
correlation
0.1
0
-0.1
-0.2
-0.3
PFC
PVC
ACC
BA
CN
C

Figure 3.

Expression levels in human cortex and cerebellum. Average expression levels (base two logarithm expression intensity; error bars indicate ± one standard error) in prefrontal cortex were calculated for four sets of genes in both young (two 45 year old) and old (one 70 year old) human samples. Red, cortex expression levels; blue, cerebellum expression levels; solid lines, genes down-regulated in frontal pole; dashed lines, genes up-regulated in frontal pole. The genes up-regulated with age in cortex are somewhat up-regulated in cerebellum, whereas those down-regulated in cortex do not change at all with age in cerebellum.

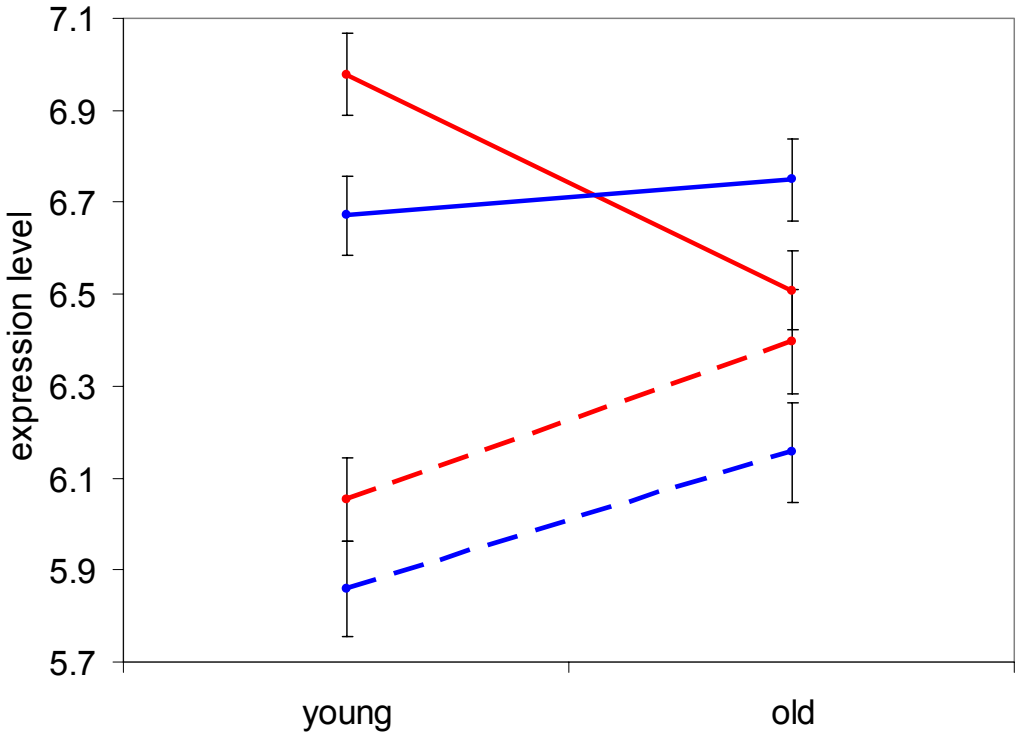
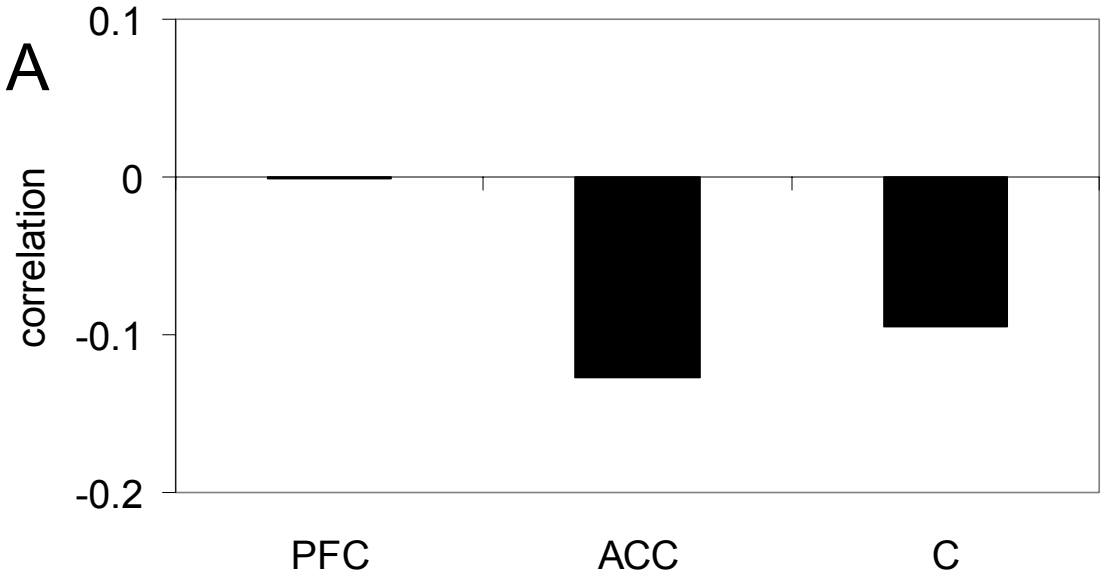

7.1
6.9
6.7
6.5
6.3
6.1
5.9
5.7
expression level
young
old

Figure 4.

Aging in the chimpanzee brain. (A) Correlations of aging gene expression patterns between human frontal pole (Lu et al 2004) and each of the three regions of the chimpanzee brain used in this work. The lack of any significant correlation indicates that human and chimpanzee brain aging patterns may differ. (B) Correlations of aging gene expression patterns between all three possible pairs of the three regions of the chimpanzee brain used in this work. The high correlation when comparing cortex regions indicates a reproducible pattern of aging in chimpanzee cortex.


A
0.1
0
-0.1
-0.2
correlation
PFC
ACC
C

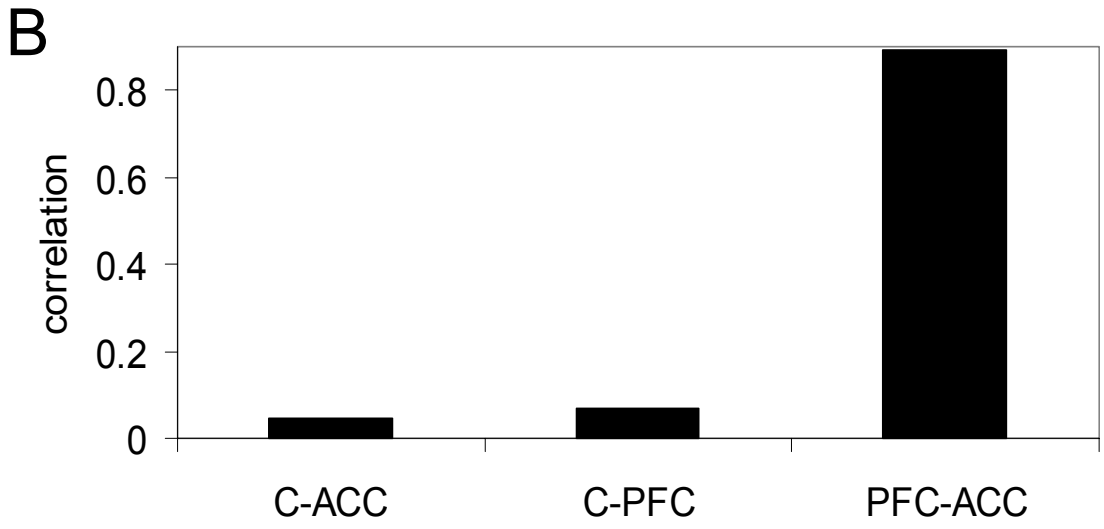
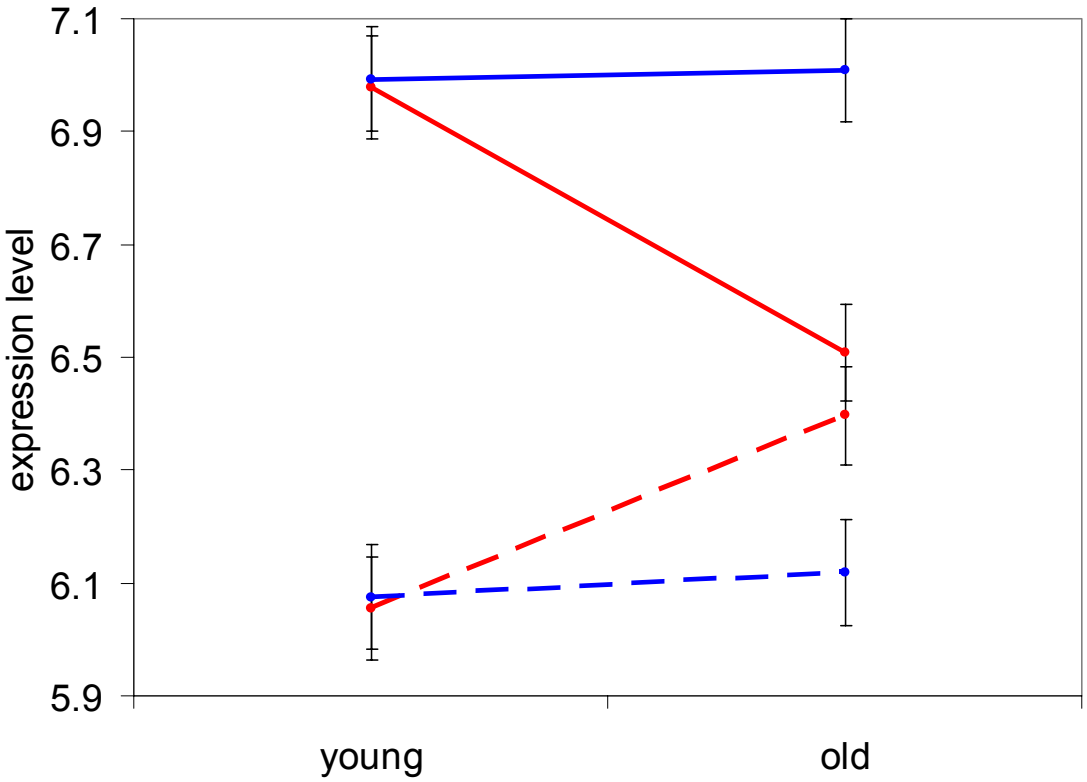

B
0.8
0.6
0.4
0.2
0
correlation
C-ACC
C-PFC
PFC-ACC

Figure 5.

Expression levels in human and chimpanzee cortex. Average expression levels (base two logarithm expression intensity; error bars indicate ± one standard error) in prefrontal cortex were calculated for four sets of genes in both young (two 45 year old human, or five 7-12 year old chimpanzee) and old (one 70 year old human, or two >40 year old chimpanzee) samples. Red, human genes; blue, chimpanzee genes; solid lines, genes (or orthologs of genes) down-regulated in human frontal pole; dashed lines, genes (or orthologs) up-regulated in human frontal pole. The chimpanzee expression levels resemble young, but not old, human.


7.1
6.9
6.7
6.5
6.3
6.1
5.9
expression level
young
old

References

Adams CC, Jakovljevic J, Roman J, Harnpicharnchai P & Woolford JL Jr. (2002) RNA 8, 150-165.

Akashi H. (2003) Genetics 164, 1291-1303.

Altschuh D, Lesk AM, Bloomer AC & Klug A (1987) J. Mol. Biol. 193, 693-707.

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO & Herschlag D (2003) Proc Natl Acad Sci U.S.A 100, 3889-3894.

Arkin A, Ross J, McAdams HH (1998). Genetics 149, 1633-1648.

Barabasi AL, Oltvai ZN (2004). Nat Rev Gen 5, 101-113.

Barkai N, Leibler S (2000) Nature 403, 267-268.

Beckman KB, Ames BN (1998). Physiological Review 78, 547-581.

Beer MA & Tavazoie S (2004) Cell 117, 185-198.

Bentourkia M, Bol A, Ivanoiu A, Labar D, Sibomana M, Coppens A, et al. (2000) . J. of Neurol. Sci. 181, 19-28.

Berg OG (1978) J. Theo. Biol. 71, 587-603.

Blake WJ, Kaern M, Cantor CR, Collins JJ (2003). Nature 422, 633-637.

Bloom JD, Adami C (2003) BMC Evolutionary Biology, 3, 21.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). Bioinformatics 19, 185-193.

Brooks PJ, Wise DS, Berry DA, Kosmoski JV, Smerdon MJ, et al. (2000). J Biol Chem. 275, 22355-22362.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, et al. (2003). Proc. Natl. Acad. Sci. USA 100, 13030-13035.

Chervitz SA et al. (1998), Science 282, 2022.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA & Johnston M (2003) Science 301, 71-76.

Corral-Debrinski M, Horton T, Lott MT, Shoffner JM, Beal MF, et al. (1992). Nature Genetics 2, 324-329.

Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, et al. (2001). Nuc Acids Res 29, 75-79.

Cutter AD, Payseur BA, Salcedo T, Estes AM, Good JM, Wood E, Hartl T, Maughan H, Strempel J, Wang B, et al. (2003) Genome Research, 13, 2651-2657.

Dandekar T, Snel B, Huynen M & Bork P (1998) Trends Biochem. Sci. 23, 324-328.

Darwin C (1859) The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. London, John Murray.

Date SV & Marcotte SM (2003) Nature Biotechnology 21, 1055-1062.

Dawkins R (1976) The Selfish Gene. Oxford: Oxford University Press.

Dickerson RE.(1971) J. Mol. Evol. 1, 26-45.

Doolittle RF (1995) Annu. Rev. Biochem. 64, 287–314.

Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, et al. (2004). Cell 199, 1027-1040.

Dragon F, et al. (2002) Nature 417, 967-970.

Eisen MB, Spellman PT, Brown PO & Botstein D (1998) Proc Natl Acad Sci U.S.A. 95, 14863-14868.

Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002). Science 297, 1183-1186.

Emerit J, Edeas M, Bricaire F (2003). Biomedicine & Pharmacotherapy 58, 39-46.

Enard W, Fassbender A, Model F, Adorjan P, Pääbo S, et al. (2004). Curr Biol 14, R148-R149.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, et al. (2002). Science 296, 340-343.

Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA (1999) Nature 402, 86-90.

Evans MD, Cooke MS (2004). Bioessays 26, 533-542.

Evans SJ, Choudary PV, Vawter MP, Li J, Meador-Woodruff JH, et al (2003). Neurobiol Dis 14, 240-250.

Feng D and Doolittle R (1997) J. Mol. Evol. 44, 361 .

Fisher RA The Genetical Theory of Natural Selection (Dover, New York, NY, 1930).

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C & Feldman MW (2002) Science 296, 750–752.

Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004). Proc Natl Acad Sci USA 101, 9033-9038.

Fraser HB, Wall DP, Hirsh AE. (2003) BMC Evolutionary Biology, 3, 11.

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Nature 415, 141-147.

Ge H, Liu Z, Church GM & Vidal M (2001) Nature Genetics 29, 482-486.

Gerhart J & Kirschner M. Cells, embryos, and evolution (Blackwell Science, Malden, MA, 1997).

Ghosh R, Mitchell DL (1999). Nucleic Acids Res. 27, 3213-3218.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002). Nature 418, 387-391.

Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP Davis RW (2004). Proc Natl Acad Sci U S A, in press.

Gibbons JD, in Sage University Papers, M. S. Lewis-Beck, Ed. (Sage Publications, Newbury Park, CA, 1993), pp. 3-29.

Goffeau A., et al. (1996) Science 274, 563-567.

Goh CS & Cohen FE (2002) J. Mol. Biol. 34, 177-192.

Goh CS, Bogan AA, Joachimiak M, Walther D & Cohen FE (2000) J. Mol. Biol. 299, 283-293.

Goldberg DS & Roth FP (2003) Proc Natl Acad Sci U.S.A. 100, 4372-4376.

Grigoriev A (2001) Nucleic Acids Res. 29, 3513-3519.

Grishin NV (1995) J. Mol. Evol. 41, 675.

Hallet B (2001). Curr Opin Microbiol 4, 570-581.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, et al (2004). Nature 430, 88-93.

Harman D (1956). J. Gerontol. 2, 298-300.

Hartwell LH, Hopfield JJ, Leibler S & Murray AW (1999) Nature 402, C47–C52.

Hekimi S, Guarente L (2003). Science 299, 1351-1354.

Hirsh AE and Fraser HB (2001) Nature 411, 1046.

Hirsh AE, Fraser HB & Wall DP (2005). Mol. Biol. Evol. 22, 174–177.

Hirsh AE, Fraser HB (2003). Nature 421, 497-498.

Ho Y et al. (2002) Nature 415, 180.

Hodgkin J (1998). Int J Dev Biol 42, 501-505.

Hoffman R,Valencia A. (2003) Trends in Genetics, 19, 681-683.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998). Cell, 95, 717-728.

Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, et al. (2000) Nature Genetics, 25, 333-337.

Ikemura T (1982) J. Mol. Biol. 158, 573–597.

Ingram VM. (1961) Nature, 189, 704-708.

Ito T et al. (2001) Proc. Natl. Acad. Sci. U.S.A. 98, 4569.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF & Gerstein M (2003) Science 302, 449-453.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001). Nature 411, 41-42.

Jordan IK, Marino-Ramirez L, Wolf YI & Koonin EV (2004) Mol. Biol. Evol. 21, 2058–2070.

Jordan IK, Wolf YI, Koonin EV. (2003) BMC Evolutionary Biology, 3, 5.

Jordan IK, Wolf YI, Koonin EV. (2003) BMC Evolutionary Biology, 3, 1.

Kaplan D, in Advanced Quantitative Techniques in the Social Sciences, J. de Leeuw, R. Berk, Eds. (Sage Publications, Thousand Oaks, CA, 2000), pp. 13-39.

Kellis M, Patterson N, Endrizzi M, Birren B & Lander ES (2003) Nature 423, 241-254.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, et al (2004). Genome Research 14, 1462-1473.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR (2000) Mol. Biol. Evol. 17, 1956.

Koski LB and Golding GB. (2001) Journal of Molecular Evolution, 52, 540-542.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003). Genome Research 13, 2229-2235.

Lee CK, Weindruch R, Prolla TA (2000). Nature Genetics 25, 294-297.

Lee TI, et al. (2002) Science 298, 799-804.

Li JZ, Vawter MP, Walsh DM, Tomita H, Evans SJ, et al (2004). Hum Mol Genet. 13, 609-616.

Lu T, Pan Y, Kao SY, Li C, Kohane I, et al (2004). Nature 429, 883-891.

Marcotte EM, Pellegrini M, Ho-Leung N, Rice DW & Yeates TO (1999) Science 285, 751-753.

Marietta C, Gulam H, Brooks PJ (2002). DNA Repair 1, 967-975.

McCarroll SA, Murphy CT, Pletcher SD, Chin CS, Jan YN, et al. (2004). Nature Genetics 36, 197-204.

Mecocci P, MacGarvey U, Kaufman AE, Koontz D, Shoffner JM, et al. (1993). Ann Neurol 34, 609-616.

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B. (2002) Nucleic Acids Res, 30, 31-34

Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y & Wang X (1994) Science 368, 251-255.

Needham, J. (1933) Biol. Rev. 8, 180–233.

Noda A, Ohba H, Kakiuchi T, Futatsubashi M, Tsukada H, et al. (2002). Brain Research 936, 76-81.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002). Nat Gen 31, 69-73.

Pal C, Papp B, Hurst LD (2003). Nature 421, 496-497.

Pal C, Papp B, Hurst LD. (2001) Genetics, 158, 927-931

Papp C, Pal B & Hurst LD (2003) Nature 424, 194-197.

Parada LA, McQueen PG, Misteli T (2004). Genome Biology 5, R44.

Partridge L, Barton NH (1993). Nature 362, 305-311.

Pazos F & Valencia A. (2001) Prot Engineering 14, 609-614.

Pazos F & Valencia A. (2002) Proteins 47, 219-227.

Pazos F, Helmer-Citterich M, Ausiello G & Valencia A. (1997) J. Mol. Biol. 271, 511-523.

Pearl R. The Rate of Living. London, Univ. of London Press, 1928.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D & Yeates TO (1999) Proc Natl Acad Sci U.S.A. 96, 4285-4288.

Poon HF, Calabrese V, Scapagnini G, Butterfield DA (2004). Clin Geriatr Med 20, 329-359.

Ptak SE, Roeder AD, Stephens M, Gilad Y, Pääbo S, et al (2004). PLoS Biology 2, e155.

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, et al. (2001) Nature, 409, 211-215

Ramani AK & Marcotte EM (2003) J. Mol. Biol. 327, 273-284.

Rawson PD, Brazeau DA, Burton RS (2000) Gene 248, 15.

Ray TS (1991) In : Belew, RK and Booker LB [eds.], Proceedings of the 1991 International Conference on Genetic Algorithms, 527-534. San Mateo, CA: Morgan Kaufmann.

Rivera MC, Jain R, Moore JE, Lake JA (1998) Proc. Natl. Acad. Sci. U.S.A. 95, 6239.

Rodwell GE, Sonu R, Zahn JM, Lund J, Wilhelmy J, et al. (2004). PLoS Biology 2, e427.

Sakaki Y, Watanabe H, Taylor T, Hattori M, Fujiyama A, et al (2003). Cold Spring Harb Symp Quant Biol 68, 455-460.

Sakamoto S, Ishii K (1999). J. of Neurol. Sci. 172, 41-48.

Schlosser G & Wagner GP (ed). Modularity in Development and Evolution (University of Chicago Press, Chicago, IL, 2004).

Schlosser G. (2002) Theory Biosci. 121, 1–80.

Schwikowski B, Uetz P, Fields S (2000) Nature Biotechnol. 18, 1257.

Sharp PM & Li WH (1987) Nucleic Acids Res. 15, 1281-1295.

Smith V, Botstein D, Brown PO (1995) Proc. Natl. Acad. Sci. U.S.A. 92, 6479.

Smith V, K., Chou N, Lashkari D, Botstein D, Brown PO (1996) Science 274, 2069.

Smolin L (1997). The Life of the Cosmos. Oxford: Oxford University Press.

Sokal RR & Rohlf FJ (1995). Biometry. New York, W.H. Freeman & Company.

Souciet JL, et al. (2000) FEBS Letters 487, 3-12.

Stadtman ER (2001). Ann N Y Acad Sci 928, 22-38.

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, et al. (2002). Nature Genetics 31, 400-404.

Swain PS, Elowitz MB, Siggia ED (2002). Proc Natl Acad Sci U S A 99, 12795-12800.

Teichmann SA (2002). J Mol Biol. 324, 399-407.

Thompson JD, Higgins DG & Gibson TJ (1994) Nucleic Acids Res. 22, 4673-4680.

Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary PV, et al (2004). Biol Psychiatry 55, 346-352.

True JR, Carroll SB (2002). Annu Rev Cell Dev Biol. 18, 53-80.

Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, et al. (2004). Proc. Natl. Acad. Sci. USA 101, 2957-2962.

Uetz P et al. (2000) Nature 403, 623.

von Dassow G, Munro E (1999). J Exp Zool 285, 307-325.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002). Nature 417, 399-403.

Wall DP, Fraser HB, and Hirsh AE. (2003) Bioinformatics 19, 1710-1711.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002). Proc Natl Acad Sci U S A 99, 5860-5865.

Waxman D & Peck JR (1998) Science 279, 1210–1213.

Welle S, Brooks A, Thornton CA (2001). Physiol Genomics 5, 67-73.

Wilson AC, Carlson SS White TJ. (1977) Ann. Rev. Biochem, 46, 573-639.

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al (2005). Science Feb 10 [Epub ahead of print].

Winzeler EA et al. (1999) Science 285, 901.

Yang AS (2001). Evol Dev 3, 59-72.

Yang Z (1997). Comput Appl Biosci 13, 555-556.

Zuckerkandl E (1976) J. Mol. Evol. 7, 167.